

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED FINAL 01 Jul 94 To 30 Jun 97	
4. TITLE AND SUBTITLE PREDICTING TOXICITY AND DEGRADABILITY OF QUADRICYCLANE, FLUOROCARBON ETHERS AND THEIR ANALOGS			5. FUNDING NUMBERS F49620-94-1-0401 2312/AS 61102F
6. AUTHOR(S) Dr Subhash C. Basak		AFOSR-TR-97-0374	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Dept of Chemical Engineering University of Minnesota, Duluth Duluth MN 55812		8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) AFOSR/NL 110 Duncan Ave Room B115 Bolling AFB DC 20332-8050 Dr Walter Kozumbo		10. SPONSORING/MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES			
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited.		12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) A number of novel molecular similarity methods have been developed using topostructural and topochemical parameters which can be computed directly from molecular structure using POLLY. Topological indices (TIs), atom pairs (APs), geometrical parameters, and semiempirical quantum chemical parameters have been used for molecular similarity analysis and development of hierarchical QSAR models. The relative effectiveness of the various similarity techniques in selecting analogs and estimating properties of toxicological importance have been tested on a selected set of properties such as mutagenicity, acute toxicity, lipophilicity (logP, octanol/water), etc. The K nearest neighbor (KNN) method, K=1,2...25, has been used in generating probe-induced subsets from different databases. Results show that the KNN method gives the best estimate of properties at K=5-10 for the properties studied.			
14. SUBJECT TERMS		15. NUMBER OF PAGES	
		16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT (U)	18. SECURITY CLASSIFICATION OF THIS PAGE (U)	19. SECURITY CLASSIFICATION OF ABSTRACT (U)	20. LIMITATION OF ABSTRACT (UL)

**Final Report
of the Air Force Project**

Covering the period 8/1/96 to 7/31/97

Agency No: DOD/F49620-94-1-0401
U of M No: 0756-5140 (1613-189-6090)
NRRI Technical Report Number: NRRI/TR-97/15

**Predicting Toxicity and Degradability of
Quadricyclane, Fluorocarbon Ethers
and Their Analogs**

Submitted By:

Subhash C. Basak, Ph.D.
Principal Investigator
Natural Resources Research Institute
5013 Miller Trunk Highway
Duluth, MN 55811
Phone: (218)720-4230 Fax: (218)720-9412
Email: sbasak@wyle.nrri.umn.edu

Keith B. Lodge, Ph.D.
Co-principal Investigator
Department of Chemical Engineering
University of Minnesota, Duluth
Duluth, MN 55812

Joseph Schubauer-Berigan, Ph.D.
Principal Investigator of the USC Subcontract
NIWB NERR
Baruch Marine Laboratory
P.O. Box 1630
Georgetown, SC 29442

The University of Minnesota is an equal opportunity educator and employer.

19971003 044

Table of Contents

Acknowledgments	2
Objectives	3
Status of Effort	3
Accomplishments/New Findings	4
TASK 1: Development of data bases	4
TASK 2: Development of methods to quantify molecular similarity	4
TASK 3 Selection of analogs	5
TASK 4	5
A. Estimation of properties of the target chemical from the probe-induced subset	
B. Hierarchical approach to toxicity estimation using topological, geometrical, and quantum chemical parameters	
Task 5 Measurement of hydrophobicity	6
Task 6 Microbial degradation studies	7
Task 7 Photochemical Degradation Studies	7
Personnel Supported	7
Publications	7
Interactions/transitions	8
A. Participation/Presentations	8
B. Consultative and Advisory Functions	9
C. Transitions	9
Honors/Awards	10
Appendix 1.....	10
Appendix 2.....	10

Acknowledgments

During the third year of the Air Force grant I have benefited through interaction with numerous colleagues. I would like to specially mention Dr. Mic Lajiness (Computer-Aided Drug Discovery, The Upjohn Company, Kalamazoo, MI), Professor A. T. Balaban (Polytechnic Institute, Bucharest, Roumania), Professor K. Balasubramanian (Department of Chemistry and Biochemistry, Arizona State University Tempe, AZ), Professor William C. Herndon, (Department of Chemistry, University of Texas, El Paso), Dr. David Opitz (Department of Computer Science, University of Minnesota, Duluth), Dr. Mark Johnson (Computer-Aided Drug Discovery, The Upjohn Company, Kalamazoo, MI), and Professor Milan Randić (Department of Mathematics and Computer Science, Drake University, IA) for many fruitful discussions on topics related to quantification of molecular structure, molecular similarity and structure-activity relationships (SAR) pertaining to environmental toxicology.

I am thankful to Mr. Greg Grunwald, Applications Programmer of NRRI, for his excellent efforts in developing databases and carrying out computations resulting in many publications supported by the Air Force.

Finally, I would like to thank AFOSR for providing us financial support for carrying out the research reported herein.



Subhash C. Basak, Ph.D.
Principal Investigator

DTIC QUALITY INSPECTED 3

Objectives (The same as in the original proposal)

In a large number of cases, we have to assess the risk of chemicals and predict the toxic potential of molecules in the face of limited experimental data. Structural criteria and functional criteria (if available) are routinely used to estimate the possible hazard posed by a chemical to the environment and ecosystem. Frequently, no biological or relevant physicochemical properties of the chemical species of interest are available to the risk assessor.

In the proposed project, we will develop and implement a number of methods of quantifying molecular similarity of chemicals using techniques of computational and mathematical chemistry. Some of the methods are new and will be based on our own research on the theoretical development and implementation of molecular similarity methods. These techniques will be implemented in a user friendly computer environment of the Silicon Graphics workstation. The similarity methods will be used to select analogs of chemicals of interest to the Air Force, *viz.*, QUADRICYCLANE, FLUOROCARBON ETHERS AND THEIR ANALOGS, from databases containing high quality physicochemical data and toxicity endpoints for large number of chemicals. The databases used in the project will come from three sources: a) public domain databases, b) our own in-house databases, and c) databases acquired from commercial vendors.

The set of selected analogs, called probe-induced subsets, will be used to: a) develop structure-activity relationships (SAR), and b) carry out ranking of chemicals. Both of these methods will be used to estimate the hazard of the chemicals of interest.

A set of chemicals (five to ten) will be chosen for experimental work with the purpose of evaluating and refining computer models. The set will include quadricyclane and fluorocarbon ethers of interest to the Air Force. It will also include a selection of analogs (probe-induced subset) that are readily available, suitable for experimentation, and for which data are lacking. Experiments will be performed to assess the biodegradability and photochemical degradability of the members of the set. Their toxicity will be tested by MicroTox and MutaTox. In cases where significant degradation is observed, the toxicity of the degradation products will also be tested. Direct measurement of the hydrophobicity (octanol-water partition coefficient) will be performed on the members of the set.

Status of Effort

A number of novel molecular similarity methods have been developed using topostructural and topochemical parameters which can be computed directly from molecular structure using POLLY. Topological indices (TIs), atom pairs (APs), geometrical parameters, and semiempirical quantum chemical parameters have been used for molecular similarity analysis and development of hierarchical QSAR models. The relative effectiveness of the various similarity techniques in selecting analogs and estimating properties of toxicological importance have been tested on a selected set of properties such as mutagenicity, acute toxicity, lipophilicity ($\log P$, octanol/water), etc. The K nearest neighbor (KNN) method, $K=1, 2, \dots, 25$, has been used in generating probe-induced subsets from different databases. Results show that the KNN method gives the best estimate of properties at $K = 5-10$ for the properties studied.

Seventy-five probe-induced subsets have been generated for Quadricyclane from three different databases: a) STARLIST $\log P$ database of Daylight, Inc., containing more than 4,000

high quality logP values, for b) the selection was restricted to C₇ cmpds Available Chemicals Directory (ACD), containing over 180,000 chemicals which are currently available from suppliers worldwide, and c) a Chemical Abstracts Service database containing about 120,000 diverse chemicals. Some of the selected analogs are being tested in the laboratory in order to determine the utility of analogs in predicting properties of chemicals from the properties of their neighbors in similarity spaces.

Accomplishments/New Findings

The following is a summary of accomplishments of the various tasks of the project during the reporting period.

TASK 1: Development of data bases

A large number of databases relevant to toxicology have been developed from published literature. These include properties like teratogenicity, inhibition of microsomal and mitochondrial oxygen uptake in rat cerebellum by chemicals, minimum inhibitory concentration of chemicals for DNA gyrase activity in E. Coli, EC₅₀ for AHH receptor activation, Ames mutagenicity, Ito's test for carcinogenicity, liver carcinogenicity in rat/mice, acute toxicity of various pollutants including pesticides, LC₅₀ in guppy, LC₅₀ in fathead minnow, skin permeability of chemicals, lowest observed adverse effect levels (LOAELs), water solubility, soil sorption coefficient, toxicity of organophosphate insecticides, and toxicity of respiratory uncouplers.

Many of these data have structural/mechanistic implications for toxicology. Some sets of compounds contain a specific toxicophore which is responsible for their particular toxic action. QSAR studies can show how the effect of the toxicophore is modulated by structural modifications. On the other hand, some toxicological data are collected on common biological endpoints of diverse structural types. These data will be used to develop similarity and hierarchical QSAR models. Mechanistic data developed by the toxicology group at the Air Force labs will be used to validate the QSAR models generated from literature data.

TASK 2: Development of methods to quantify molecular similarity

New molecular similarity methods have been developed using topostructural indices, topochemical parameters, atom pairs (APs) and geometrical parameters. A hierarchical approach to the quantification of molecular similarity has been developed in a limited scale.

Principal components analysis (PCA) and variable clustering methods have been used to create orthogonal structure spaces from POLLY parameters. AP based similarity methods have also been compared with PCA based methods in the selection of analogs and prediction of properties.

The following publications reported the result of molecular similarity analysis:

- 1) Use of graph-theoretic parameters in predicting inhibition of microsomal p-hydroxylation of aniline by alcohols: a molecular similarity approach by Subhash C. Basak and Brian D. Gute., pp. 492-504, *In: Proceedings of the 2nd International Congress on Hazardous Waste: Impacts on Human and Ecological Health*, B.L. Johnson, C. Xintaras, J.S. Andrews, Jr., Eds., Princeton Scientific Publishing Co. Inc., Princeton, New Jersey. 1997.

This paper compares the relative effectiveness of the Euclidean distance (ED) and AP methods in estimating the inhibition of microsomal p-hydroxylation of aniline by alcohols.

2) Estimation of the normal boiling points of haloalkanes using molecular similarity by Subhash C. Basak, Brian D. Gute, and Gregory D. Grunwald. *Croatia Chemica Acta* **69**:1159-1173, 1996.
This paper estimated the normal boiling points of a set of 267 haloalkanes using molecular similarity methods.

3) Use of graph-theoretic and geometrical molecular descriptors in structure-activity relationships, by Subhash C. Basak, Gregory D. Grunwald, and Gerald J. Niemi. pp. 73-116, *In: From Chemical Topology to Three-Dimensional Geometry*, ed. A.T. Balaban, Plenum Press, New York, 1997.
This book chapter presents a comprehensive review of the utility of topological indices in QSAR and the quantification of intermolecular similarity.

4) Characterization of the molecular similarity of chemicals using topological invariants, by Subhash C. Basak, Brian D. Gute, and Gregory D. Grunwald. *In: Advances in Molecular Similarity*, JAI Press, submitted, 1997.
This paper analyzed the utility of topostructural and topochemical indices in the quantification of molecular similarity and selection of analogs.

Copies of the above mentioned papers are attached (Vide Infra, Publication Section)

TASK 3 Selection of analogs

Analogs or "probe-induced subsets" selected from databases with good quality experimental data can be useful in predicting properties of probe chemicals. Taking Quadricyclane as the probe, we selected 75 analogs using different search methods. The results of such analyses have been previously reported. Various molecular similarity methods have also been used in the selection of neighbors for KNN based property estimation.

TASK 4

A. Estimation of properties of the target chemical from the probe-induced subset
We studied the effectiveness of similarity methods developed in Task 2 above by applying these methods in estimating various physicochemical and toxicological endpoints. To this end, we carried out similarity-based estimation of physicochemical and toxicological properties. Three papers were submitted out of the research carried out in this task. These results were also presented in numerous national and international symposia and invited presentations.

- a. Use of graph-theoretic parameters in predicting inhibition of microsomal p-hydroxylation of aniline by alcohols: a molecular similarity approach, by Subhash C. Basak and Brian D. Gute., pp. 492-504, *In: Proceedings of the 2nd International Congress on Hazardous Waste: Impacts on Human and Ecological Health*, B.L. Johnson, C. Xintaras, J.S. Andrews, Jr., Eds., Princeton Scientific Publishing Co. Inc., Princeton, New Jersey, 1997.
- b. Estimation of the normal boiling points of haloalkanes using molecular similarity by Subhash C. Basak, Brian D. Gute, and Gregory D. Grunwald. *Croatia Chemica Acta*, **69**:1159-1173, 1996.
- c. Characterization of the molecular similarity of chemicals using topological invariants, by Subhash C. Basak, Brian D. Gute, and Gregory D. Grunwald., *In: Advances in Molecular Similarity*, JAI Press, submitted, 1997.

B. Hierarchical approach to toxicity estimation using topological, geometrical, and quantum chemical parameters

We have also been developing a hierarchical approach to computational toxicology using topostructural, topochemical, geometrical, and quantum chemical parameters which can be calculated directly from molecular structure. This approach uses increasingly more complex parameters to estimate properties of chemicals, as necessary for a particular situation. We have listed below the book chapters/papers in peer-reviewed journals which have reported these results.

- 5) A comparative study of topological and geometrical parameters in estimating normal boiling point and octanol/water partition coefficient, Subhash C. Basak, Brian D. Gute, and Gregory D. Grunwald. *J. Chem. Inf. Comput. Sci.* **36**:1054-1060, 1996.
This paper used topostructural, topochemical and geometrical parameters in the development of hierarchical QSAR models for predicting logP (octanol/water) and boiling point.
- 6) Use of topostructural, topochemical and geometric parameters in the prediction of vapor pressure: a hierarchical QSAR approach, S. C. Basak, B. D. Gute and G. D. Grunwald, *J. Chem. Inf. Comput. Sci.*, **37**: 651-655, 1997.
This paper utilized a hierarchical QSAR approach in estimating vapor pressure of a diverse set of 476 chemicals.
- 7) Predicting acute toxicity (LC₅₀) of benzene derivatives using theoretical molecular descriptors: a hierarchical QSAR approach, B. D. Gute and S. C. Basak, *SAR QSAR Environ. Res.*, in press, 1997.
This paper used a hierarchical QSAR approach in the estimating acute toxicity of a set 69 of benzene derivatives. Topostructural, topochemical, geometrical and quantum chemical parameters were used as independent variables.
- 8) Characterization of molecular structures using topological indices, S.C. Basak and B.D. Gute; *SAR QSAR Environ. Res.*, in press, 1997.
- 9) The relative effectiveness of topological, geometrical, and quantum chemical parameters in estimating mutagenicity of chemicals, S. C. Basak, B. D. Gute and G. D. Grunwald. *In: Proceedings of the Seventh International Workshop on Quantitative Structure-Activity Relationships in Environmental Sciences*, SETAC Press, in press, 1997.
This paper used a hierarchical approach in estimating mutagenicity of chemicals.

Copies of the above papers are attached (Vide Infra, Publication Section)

Task 5 Measurement of hydrophobicity

We measured the octanol-water partition coefficient (P) for 15 analogs. Because the application of the retention-time method (see the Annual Report for Year 2) gave values of logP about an order-of-magnitude greater than those predicted by CLOGP, we considered it worthwhile to do the measurements thoroughly. The results are shown in Table 1. For eight of the compounds we used both a stirring method and a shake-flask method; the results from both methods agree well. We filled a gap around logP = 2 with the compound 2-norbornane methanol (a mixture of exo and endo). Figure 1 gives a plot of measured logP vs estimated logP (CLOGP) for the analogs tested.

Task 6 Microbial degradation studies

Biodegradation and toxicity of quadricyclane and its six analogs selected by molecular similarity methods have been determined (See Appendix 1 for details). Results indicate that both quadricyclane and its selected analogs are readily degradable.

Task 7 Photochemical Degradation Studies

We carried out photochemical reactions with hydrogen peroxide on 6 additional compounds, *viz.* endo-norborneol, exo-norborneol, 3,5-dihydroxytricyclo[2.2.1.0]heptane, 2,7-norbornanediol, dicyclopentylcarbinol and *cis*-*exo*-2,3-norbornanediol. (The experimental details are given in the Annual Report for Year 2). We observed no significant reactions in these cases; the data are not shown.

Personnel Supported

University of Minnesota, Duluth

Subhash Basak, Keith Lodge, Greg
Grunwald, Gloria Bly, and A. Hayford

University of South Carolina

Joseph Schubauer-Berigan, and Darcy
Wood

Publications

The following publications, which are currently either published, accepted for publication or in submission, report results of QSAR/QMSA analyses which were supported by the AFOSR grant:

1. Use of graph-theoretic parameters in predicting inhibition of microsomal p-hydroxylation of aniline by alcohols: a molecular similarity approach, Subhash C. Basak and Brian D. Gute., pp. 492-504, *In: Proceedings of the 2nd International Congress on Hazardous Waste: Impacts on Human and Ecological Health*, B.L. Johnson, C. Xintaras, J.S. Andrews, Jr., Eds., Princeton Scientific Publishing Co. Inc., Princeton, New Jersey, 1997.
2. Estimation of the normal boiling points of haloalkanes using molecular similarity, Subhash C. Basak, Brian D. Gute, and Gregory D. Grunwald. *Croatia Chemica Acta*, **69**:1159-1173, 1996.
3. Use of graph-theoretic and geometrical molecular descriptors in structure-activity relationships, Subhash C. Basak, Gregory D. Grunwald, and Gerald J. Niemi. pp. 73-116, *In: From Chemical Topology to Three-Dimensional Geometry*, ed. A.T. Balaban, Plenum Press, New York, 1997.
4. Characterization of the molecular similarity of chemicals using topological invariants, S. C. Basak, B. D. Gute, and G. D. Grunwald, *In: Advances in Molecular Similarity*, JAI Press, submitted, 1997.
5. A comparative study of topological and geometrical parameters in estimating normal boiling point and octanol/water partition coefficient, Subhash C. Basak, Brian D. Gute, and Gregory D. Grunwald. *J. Chem. Inf. Comput. Sci.*, **36**:1054-1060, 1996.

6. Use of topostructural, topochemical and geometric parameters in the prediction of vapor pressure: a hierarchical QSAR approach, S. C. Basak, B. D. Gute and G. D. Grunwald. *J. Chem. Inf. Comput. Sci.*, **37**:651-655, 1997.
7. Predicting acute toxicity (LC₅₀) of benzene derivatives using theoretical molecular descriptors: a hierarchical QSAR approach, B. D. Gute and S. C. Basak. *SAR QSAR Environ. Res.*, in press, 1997.
8. Characterization of molecular structures using topological indices, S.C. Basak and B.D. Gute. *SAR QSAR Environ. Res.*, in press, 1997.
9. The relative effectiveness of topological, geometrical, and quantum chemical parameters in estimating mutagenicity of chemicals, S. C. Basak, B. D. Gute and G. D. Grunwald. *In: Proceedings of the Seventh International Workshop on Quantitative Structure-Activity Relationships in Environmental Sciences*, SETAC Press, in press, 1997.
10. On the relationship between the organic-carbon normalized sediment, or soil sorption coefficient and the octanol-water partition coefficient. K. B. Lodge. Res. Notes, submitted, 1997.

Interactions/transitions

A. Participation/Presentations

1. Subhash C. Basak and Brian D. Gute presented an invited lecture at the international symposium organized for the 1995 Herman Skolnick award in chemical information. The symposium was part of the American Chemical Society meeting, Orlando, Florida, August 25-29, 1996.
2. Subhash C. Basak and Brian D. Gute gave an invited presentation "Quantitative Molecular Similarity Analysis (QMSA) and Toxicity Prediction" at the US Air Force Conference "Chemistry and Toxicology of Candidate Deicers" organized by the Materials Directorate of Wright Patterson Air Force Base (WPAFB), Dayton, OH. While there, Dr. Basak also attended the Air Force Office of Scientific Research (AFOSR) Dermal Focus Group Meeting organized at WPAFB, August 6-7, 1996.
3. Subhash C. Basak presented a seminar "QSAR/QMSA using nonempirical parameters: applications in predictive toxicology and drug discovery" at the Abbott Laboratories, Chicago, September 22-23, 1996.
4. Brian D. Gute, Subhash C. Basak and Greg D. Grunwald gave a presentation entitled "Development of QSARs of bioactive molecules using a hierarchical approach" at the American Chemical Society 31st Midwest Regional meeting, November 6-8, 1996.
5. Subhash C. Basak gave a presentation entitled "Development of QMSA and QSAR methods for hazard assessment of chemicals: tools for computational toxicology" at the Air Force Office of Scientific Research (AFOSR) Toxicology Program Review, December 12-13, 1996, Fairborn, Ohio.

6. Subhash C. Basak presented a seminar "Computational chemical graph theory and its practical applications" in the Scientific Computing Seminar Laboratory for Intelligent Systems - ECE Dept. and CSc Dept., University of Minnesota, Duluth on January 29, 1997.
7. Subhash C. Basak, Brian D. Gute and Greg D. Grunwald presented an invited paper entitled "Use of nonempirical structural descriptors in QSAR" in the session "Mathematical approaches to QSAR and predictive toxicology" of the 11th International Conference on Mathematical and Computer Modelling and Scientific Computing in Washington, DC, March 27-April 3, 1997.
8. Subhash C. Basak, Brian D. Gute, and Greg D. Grunwald presented an invited paper entitled "Use of theoretical molecular descriptors in structure-property and structure-activity studies" at the 7th International Conference on Mathematical Chemistry and 3rd Girona Seminar on Molecular Similarity, Girona, Spain, May 26-31, 1997.
9. Subhash C. Basak presented an invited lecture entitled "Prediction of physicochemical and toxicological properties of chemicals using theoretical molecular descriptors" at Moscow State University, Moscow, Russia, June 30, 1997.
10. Subhash C. Basak, Brian D. Gute, and Greg D. Grunwald gave an invited lecture entitled "Predicting bioactivity of chemicals from structure: a hierarchical QSAR approach" to the Department of Biochemistry, University of Calcutta, Calcutta, India, July 30, 1997.

B. Consultative and Advisory Functions

Subhash C. Basak was invited to become a member of the National Advisory Board of the Association of Ayurvedic Doctors of India (AADI).

C. Transitions

1. Computational methods were applied in the design of new anti-epileptic compounds in cooperation with Professor Alexandru T. Balaban, Vice President, Rumanian Academy of Sciences.
2. Applied similarity and QSAR methods in the design of novel and benign deicing agents working in cooperation with Professor George Mushrush, Department of Chemistry, George Mason University, Washington D.C.

New Discoveries

1. Hierarchical QSAR research using topostructural, topochemical, and geometrical parameters showed that the first two classes of parameters explain most of the variance in the data of toxicological and physicochemical properties.
2. It was observed that similarity spaces derived from topostructural and topochemical parameters have distinct analog selection characteristics.

Honors/Awards

1. Subhash C. Basak was invited to become one of six invited speakers at the international symposium organized for the 1995 Herman Skolnick Award in Chemical Information. The symposium was held during the American Chemical Society Meeting, Orlando, Florida, August 25-29, 1996, to honor Milan Randić, the recipient of 1995 Herman Skolnick Award.
2. Subhash C. Basak was invited to chair and organize two sessions at the 11th International Conference on Mathematical and Modelling and Scientific Computing, March 31-April 3, 1997, Georgetown University, Washington, DC.
3. Subhash C. Basak was invited to edit a special volume of the journal *Mathematical Modelling and Scientific Computing* dealing with the mathematical aspects of QSAR and predictive toxicology.
4. Subhash C. Basak was invited to become a member of the Organizing and Scientific Committee for the International Conference on Mathematical and Computer Modelling and Scientific Computing.
5. Subhash C. Basak was invited to present a lecture on molecular similarity at the 7th International Conference on Mathematical Chemistry and 3rd Girona Seminar on Molecular Similarity, Girona, Spain, May 26-31, 1997.
6. Subhash C. Basak has been invited to deliver a plenary lecture at the 17th Annual Convention of the Indian Association for Cancer Research and National Symposium on Breast Cancer to be held in Calcutta, India, January 21-24, 1998.

APPENDIX 1.

Annual progress report of the University of South Carolina subcontract for the AFOSR grant F 49620-94-1401

APPENDIX 2.

Publications.

Table I. Direct Measurements of Log P

Compound	Shake-Flask Method				Stirring Method				Overall Sample Standard Deviation in Log P	Number of Determinations, N	CLOGP	Note Number
	Analog No.	Log P	Sample Standard Deviation in Log P	Number of Determinations	Sample Standard Deviation in Log P	Number of Determinations	Log P					
3,5-dihydroxytricyclo[2.2.1.0(2,6)]heptane	1	0.40	0.02	4	-0.42	0.05	3	-0.42	0.05	3	-1.87	
2,7-norbornanediol	11							0.40	0.02	4	-1.07	
analog #10	10										-0.21	1,5
cis-exo-2,3-norbornanediol	12	0.47	0.01	4	0.998	0.002	3	0.994	0.004	7	0.54	2
5-norbornen-2-ol (endo)	30	0.991	0.004	4	1.242	0.006	3	1.243	0.007	7	0.54	2
5-norbornen-2-ol (exo)	30	1.243	0.009	4				1.07	0.01	4	0.96	
dicyclopropyl carbinal	16	1.07	0.01	4	0.81	0.01	3	0.81	0.01	7	-0.18	
nortricyclene-3	4	0.80	0.01	4	1.39	0.07	3	1.37	0.06	5	1.02	
(+/-)-exo-norbornol	18	1.35	0.07	2				1.45	0.01	4	1.02	
(+/-)-endo-norbornol	17	1.45	0.01	4				1.99	0.01	4	1.64	3
2-norbornane methanol (exo & endo)					1.99	0.01	4					
exo-2,3-epoxynorbornane	5	1.575	0.001	2	1.580	0.006	3	1.578	0.005	5	0.40	
bicyclohepta-2,5-diene	80	2.67	0.01	2	2.67	0.01	3	2.67	0.01	5	2.11	
quadricyclane	0	2.90	0.03	8				2.90	0.03	8	1.50	
nortricyclyl bromide	2										2.28	4,5
norbornylene	60	3.24	0.01	2	3.25	0.01	3	3.24	0.01	5	2.63	
(+/-)-exo-2-chloronorbornane	15										2.94	4,5
exo-2-bromonorbornane	14	3.79	0.02	4	3.78	0.04	3	3.78	0.03	7	3.08	4,5
norbornane	23											

Table II. Indirect Measurements of Log P

Compound	Analog No.	Log P	Standard Error	CLOGP
analog #10	10	0.4	0.2	-0.21
nortricyclyl bromide	2	3.1	0.2	2.28
(+/-)-exo-2-chloronorbornane	15	3.3	0.2	2.94
exo-2-bromonorbornane	14	3.6	0.2	3.08

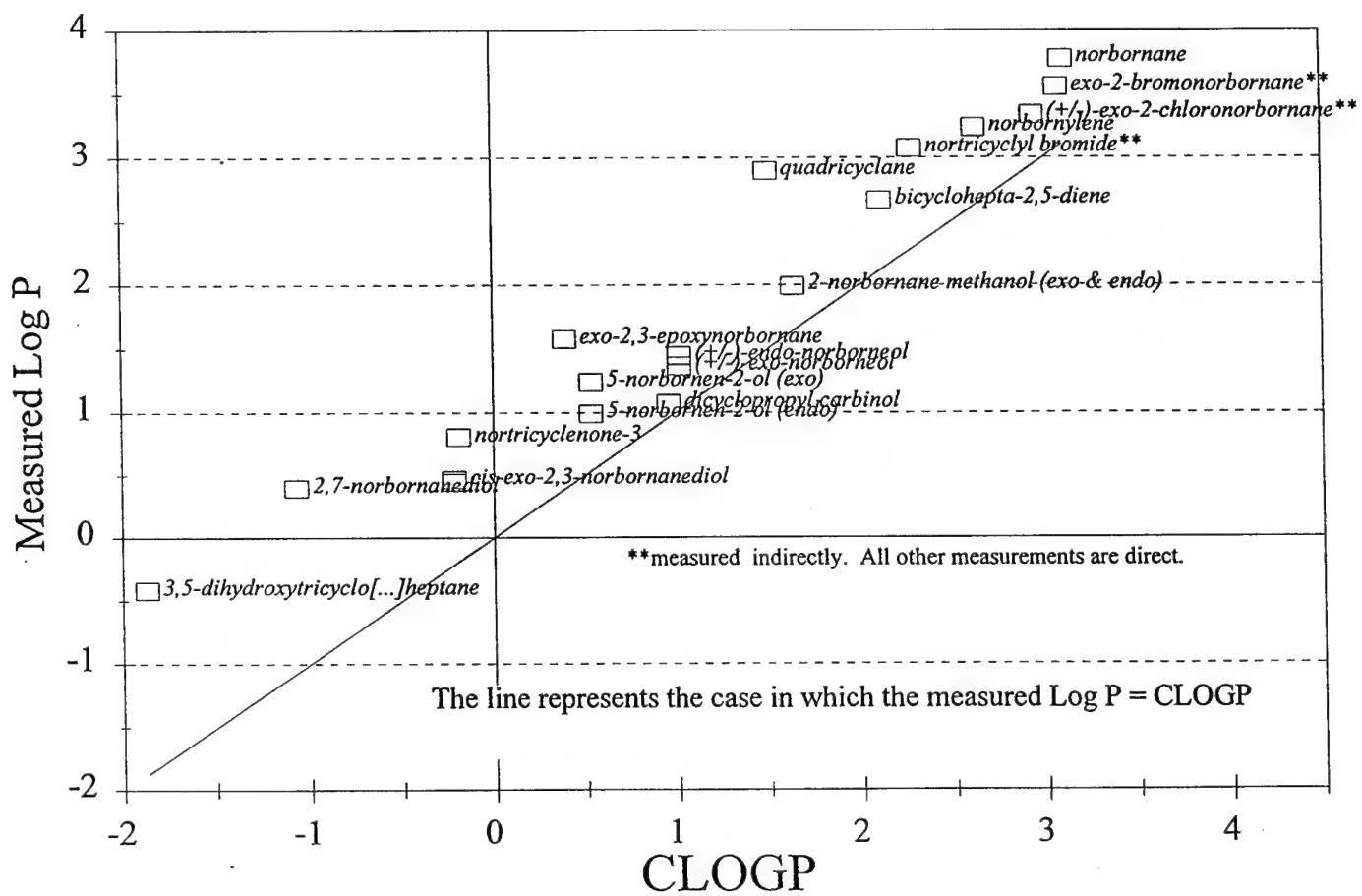
General Notes:

In cases in which both the stirring method and the shake-flask method were used, the approach to equilibrium was monitored for both methods. In this way we are confident that equilibrium is attained for those cases in which the shake-flask method alone (with no monitoring of the approach to equilibrium) is used. The values of CLOGP are calculated with release 4.51 of the Daylight Software.

Specific Notes:

- Analog 10 is 2,3-norbornanediol; the configuration of the hydroxy groups is uncertain (cis-endo or trans?) -- see next compound. 2. The assignment of "endo" and "exo" needs confirmation -- compare the results for exo- and endo-norbornol.
- 2-norbornane methanol (exo and endo) was added to the set; it fills in a gap around Log P = 2. 4. The halogenated compounds hydrolyse rapidly; this makes direct measurements impractical. 5. Estimates of measured Log P values are provided here using the retention time method (Indirect Method) in which the calibration compounds are those for which Log P was measured directly here.

Figure. Measurements of the octanol-water partition coefficient



Annual Progress of Report for AFOSR/PKC Grant F49620-94-1-0401

To the University of Minnesota

Subcontract 1613-189-6090-7901 To The University of South Carolina

Submitted By Dr. Joseph P. Schubauer-Berigan

August 24, 1997

Period Covered: (July 1, 1996 - August 24, 1997)

Task 6 Microbial degradation studies

Progress this year In past year we have primarily focused on the quantifying the biodegradation and toxicity of ODC and 6 of its analogs (identified by the work of Dr. Basak). In correspondence to the AFOSR in March and May 1997 we requested a 3 month no cost project extension and permission to reprogram some of our remaining grant funds to buy a liquid autosampler and computer for the gas chromatograph to expedite sample analysis. Both requests were granted in June. The equipment was purchased in June and in place and functioning by the end of July 1997. This equipment has allowed us to complete a substantial portion of the biodegradation work which is a major objective of the grant.

Biodegradation experiments We have finished examining the aerobic biodegradation of all seven compounds in fresh water wetland sediments and water, in live and dead incubations. The degradation of each chemical was examined in separate time course experiments each run for 33 days. As we reported earlier, we went to great lengths to characterize the sediments used in these experiments for a variety of parameters that could influence the rates of chemical degradation. Preliminary analysis of the experiments suggests very similar degradation rates among the chemicals, but analysis of the data from the experiments is incomplete. In September 1997 we will be completing the analysis of the experiments. We also hope to complete a series of experiments examining the anaerobic degradation of QDC in fresh water and sediments and the degradation of QDC in saline water and sediments.

Toxicity determinations- We have already assessed the toxicity of QDC and its analogs using BOD's and natural bacterial CFU's. This past year we developed a new and novel approach to assess the toxicity and environmental risk of these chemicals to natural bacterial community function utilizing BIOLOG plates. Briefly, the method involves directly incubating natural water samples titrated with the chemical of interest in BIOLOG plates. The plates contain 95 different carbon substrates. We monitor the resulting community-dependent substrate utilization patterns by the irreversible reduction of a tetrazolium dye associated with each substrate. In this way we can quantify the effect or toxicity of a chemical on the metabolism of classes of substrates by natural microbial communities. By examining the intensity of the dye reduction over a period of days we can also assess the effect the chemical has on the average rate of substrate metabolism by the microbial community. This approach is exciting because it give us a way to rapidly, directly and specifically assess the risk associated with these chemicals to natural ecosystem function. In comparison, one of the other approaches we are using to assess the toxicity of the chemicals is Microtox. Although this approach works well in comparing the relative toxicity of QDC and its analogs and potentially other

chemicals, one of the major weaknesses of Microtox is its inability to directly relate the results to environmental risk in the field.

We have now completed tests of all of the chemicals using the modified BIOLOG approach we developed, but analysis of the data is not complete. The toxicity assessment of the chemicals using the Microtox method is underway but not completed. Preliminary results of modified BIOLOG tests suggest widespread disruption of natural microbial community metabolism at concentrations greater than approximately 250-300 mg/L for QDC and its analogs. Preliminary Microtox assays also indicate toxicity at similar concentrations. In comparison, previous experiments showed no clear effect of the chemicals on BOD's or bacterial species or numbers (CFU's and direct counts).

Problems encountered One continuing experimental problem we are having is directly related to the relatively low water solubility of the class of chemicals we are examining. This has prevented us from keeping the chemicals in solution at the highest test levels (>300-400 mg/l). We have partially gotten around this by using ethanol to carry the chemicals. However at the highest concentrations we are also limited by the toxicity of the carrier solvents themselves which we have empirically determined.

Future plans In the time remaining on the grant, we plan to: 1) finish the Microtox toxicity tests; 2) finish the biodegradation studies mentioned above; and 3) finish analyzing the data from the toxicity and biodegradation studies. We expect at least three publications to come from the Microbial experiments: 1) a synthesis paper with the other PI's of the study; 2) a paper describing the degradation patterns of QDC and the analogs authored jointly with Dr. Lodge; and 3) a paper comparing toxicity of QDC and its analogs, including a description of our modified BIOLOG approach to assess chemical risk.

4) Accomplishments/New Findings: The research completed thus far indicates that Quadricyclane and its analogs (selected using QSAR-SAR methods) all appear to degrade rapidly, primarily abiotically, in water and sediment. Toxicity of QDC and its analogs to natural microbial community function was noted using a newly developed assessment method.

5) Personnel Supported: Dr. Joseph P. Schubauer-Berigan, Project Principal Investigator, 22% effort; Ms. Darcy Wood, project technician, 100% effort.

6) Publications: None during this period.

7) Interactions/Transactions: None during this period.

8) New discoveries, inventions or patent disclosures: None during this period.

9) Honors and Awards:

R.A. Sheldon Scholarship, University of Georgia, 1986

Research Internship, University of Georgia Marine Institute, 1986

Regents Award for Outstanding Teaching and Research, U. GA., 1985

University-wide Fellowships, University of Georgia, 1984,86,87

Research Fellowship, Savannah River Ecology Laboratory, SC, 1977-78

NSF/AEC Research Internship, Savannah River Ecology Laboratory, SC, 1976

USE OF GRAPH-THEORETIC PARAMETERS IN PREDICTING INHIBITION OF MICROSOMAL P-HYDROXYLATION OF ANILINE BY ALCOHOLS: A MOLECULAR SIMILARITY APPROACH¹

Subhash C Basak, Brian D Gute, Natural Resources Research Institute, University of Minnesota, Duluth

INTRODUCTION

Environmental and human health risk assessment of chemicals is often carried out using insufficient experimental data. This is true for the large number of industrial chemicals, as well as for substances identified in industrial effluent, hazardous waste sites and environmental monitoring surveys (Auer et al. 1990). In 1984, the National Research Council studied the availability of toxicity data on industrial chemicals, and found that many of these chemicals have little or no test data (NRC 1984). About 13 million distinct chemicals have been registered with Chemical Abstract Service (CAS), and the list is growing by nearly 500,000 per year. Out of these chemicals, about 1,000 enter into societal use every year (Arcos 1987). Few of these chemicals are submitted with the empirical data necessary for risk assessment. In the United States, the Toxic Substances Control Act (TSCA) Inventory has over 74,000 entries, and the list is growing by nearly 3,000 per year (Auer et al. 1990, TSCA 1976). Of the approximately 3,000 chemicals submitted yearly to the United States Environmental Protection Agency (EPA) for the premanufacture notification (PMN) process, more than 50% have no experimental data at all, less than 15% have empirical mutagenicity data, and only about 6% have experimental ecotoxicological and environmental fate data. This dearth of empirical data is also true for many of the over 700 chemicals found on the Superfund list of hazardous substances (Auer et al. 1990).

A large number of physicochemical and biological test data on chemicals are a prerequisite to the proper estimation of the hazards posed by a chemical species. Table 1 gives a partial list of such properties. As a result of this lack of relevant data, a variety of structural, physicochemical, and biochemical properties are used in hazard estimation. For example, in assessing the carcinogenic potential of chemicals, three classes of criteria have been used by experts:

- Structural
- Functional
- Guilt by association

¹ This paper is contribution number 154 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported, in part, by grant F49620-94-1-0401 from the United States Air Force, Cooperative Agreement CR-819621 from the United States Environmental Protection Agency, Exxon Biomedical Sciences, Inc. and the Structure-Activity Relationship Consortium (SARCON) of the Natural Resources Research Institute at the University of Minnesota. The authors would also like to extend their thanks for Greg Grunwald's helpful discussions.

Structural criteria consist of structural analogy of a chemical species with known and well-established chemical carcinogens. Structural factors considered could be molecular size, shape, branching pattern, symmetry, and charge, to list just a few. Frequently, structural characteristics of molecules are not enough for reliable estimation of carcinogenic risk. Functional criteria, that is, results of short-term tests, may often supplement structural criteria to more reliably estimate the hazard. Ames test, mammalian cell transformation, and unscheduled DNA synthesis or pattern of regulation of gene expression are examples of frequently used functional criteria relevant to carcinogenic risk assessment. The guilt-by-association criteria consider that, even though a chemical may have been found to be inactive through normal testing (bioassays), it may need to be retested more stringently if it belongs to a class of compounds which contains potent carcinogens (Arcos 1987).

In the assessment of environmental hazards, often the physicochemical and biological test data essential to hazard estimation are unavailable. In such cases, regulators use a two-tiered approach to predict hazard from chemical structure: class-specific quantitative structure-activity relationship (QSAR) models and chemical analogs (Auer et al. 1990).

Table 1. List of properties necessary for risk assessment of chemicals.

Physicochemical	Biological
Molar Volume	Receptor Binding (K_D)
Boiling Point	Michaelis Constant (K_m)
Melting Point	Inhibitor Constant (K_i)
Vapor Pressure	Biodegradation
Aqueous Solubility	Bioconcentration
Dissociation Constant (pK_a)	Alkylation Profile
Partition Coefficient	Metabolic Profile
:Octanol-water ($\log P$)	Chronic Toxicity
:Air-Water	Carcinogenicity
:Sediment-Water	Mutagenicity
Reactivity (Electrophile)	Acute Toxicity
	: LD_{50}
	: LC_{50}

QSARs are mathematical models that use various quantifiers of chemical structure and empirical parameters (or properties) in predicting physicochemical and biological properties of molecules (Basak et al. 1990, Hansch 1976). In class-specific QSARs, a chemical is first assigned a specific structural class and the QSAR of that particular class of chemicals is used to predict the potential toxicity of the molecule of interest.

If a chemical is very complex, that is, contains many functional groups, a simplistic attempt at classification is almost certain to fail. The use of class-specific QSARs in hazard assessment of such chemicals will be limited. In such cases, one resorts to the approach of selecting analogs of the chemical of interest and estimating the hazardous potential of the chemical from the toxicity of its analogs. Analogs of new chemicals are routinely used by regulatory agencies like EPA in hazard assessment (Auer et al. 1990). Chemical X is considered to be an analog of (or similar to) chemical Y if X and Y resemble each other in one or more critical aspects, that is, structurally, stereo-electronically, or physicochemically. The use of analogs is based on the tacit assumption that similar structures have similar properties (Johnson et al. 1988, Johnson and Maggiora 1990).

A perusal of approaches used in carcinogenic risk assessment and in environmental and ecotoxicological hazard estimation indicates that the candidate chemical is compared with known toxicants, using structural or functional criteria. Experts often select these analogs (structurally related chemicals) based on their individual judgements and a selected set of structural features.

Chemical analogs can be selected using empirical descriptors or theoretical molecular descriptors (Basak and Grunwald, In Press). The paucity of available experimental data for environmental pollutants makes it desirable to develop methods for selecting chemical analogs, using nonempirical variables, which are computed directly from molecular structure (Basak and Grunwald, In Press).

In recent years, we have developed several methods for quantifying intermolecular similarity. Such methods are based on topological indices (TIs) and substructural variables, like atom pairs. TIs are numerical graph invariants that encode information like size, shape, branching pattern, symmetry and certain aspects of stereo-electronic factors associated with molecules (see TI symbols and definitions in Table 3.) Topological parameters can be useful in predicting physicochemical as well as biological properties of many different congeneric sets of molecules (Basak 1988; Basak and Grunwald 1993; Basak et al. 1982, 1983, 1984, 1986, 1987a, 1987b, 1990, 1991; Kier and Hall 1986; Niemi et al. 1992; Randić 1975). Molecular similarity methods based on substructures and TIs, have been used successfully in selecting analogs, in discovering novel drugs active against human immunodeficiency virus (HIV), and in estimating different physicochemical and toxicological properties (Basak et al. 1988, 1994, In Press b; Basak and Grunwald 1994, 1995a, 1995b, 1995c, 1995d, In Press a; Lajiness 1990; Wilkins and Randić 1980).

In this paper, similarity methods based on topological indices and atom pairs have been used to estimate the inhibitory effects (pIC_{50}) of a set of 19 aliphatic alcohols on microsomal p -hydroxylation of anilines by cytochrome P_{450} .

DATABASE

Experimental pIC_{50} values for inhibition of microsomal cytochrome P_{450} p -hydroxylation of anilines by nineteen alkanols are in Table 2 (Cohen and Mannering 1973). The original set contained 20 compounds; one, methanol, was deleted. Because of its single, unique atom pair, similarity of methanol to other compounds cannot be computed using the atom pair method.

COMPUTATION OF PARAMETERS

Topological Indices

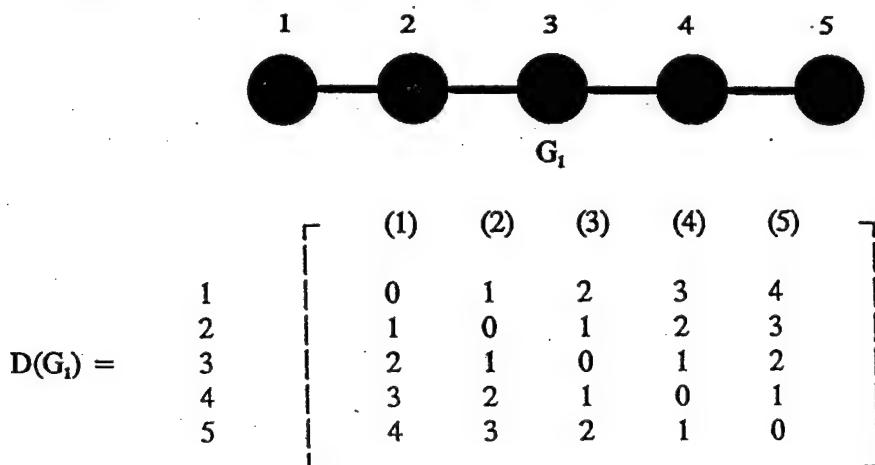
The 64 TIs in this study were calculated using POLLY 2.3 (Table 4), which can calculate 98 TIs from SMILES line notation input of chemical structures (Basak et al. 1988a). TIs include Wiener index (Wiener 1947), connectivity indices (Kier and Hall 1986, Randić 1975), information theoretic indices defined on distance matrices of graphs (Bonchev and Trinajstić 1977, Raychaudhury et al. 1984), parameters derived on the neighborhood complexity of vertices in hydrogen-filled molecular graphs (Basak 1987, Basak and Magnuson 1983, Basak et al. 1980, Roy et al. 1984), path lengths, and Balaban's J indices (1982, 1983, 1986).

Methods for calculating a few TIs used in this paper follow. The Wiener index W , the first topological index reported in the chemical literature, may be calculated from the distance matrix $D(G)$ of a hydrogen-suppressed chemical graph G as the sum of the entries in the upper triangular distance submatrix. The distance matrix $D(G)$ of a nondirected graph G with n vertices is a symmetric $n \times n$ matrix with elements d_{ij} equal to the distance between vertices v_i and v_j in G . Each

Table 2. 19 Alkanols and their observed and predicted inhibition of microsomal p-hydroxylation of anilines (pIC_{50}), by atom pair (AP) and Euclidean distance (ED) methods

Alkanol	obs. pIC_{50}	est. pIC_{50} — AP method					est. pIC_{50} — ED method				
		1	2	3	4	5	1	2	3	4	5
Ethanol	-1.10	-0.48	-0.48	-0.34	-0.34	-0.15	-0.48	-0.48	-0.33	-0.34	-0.22
1-Propanol	-0.48	-0.05	-0.05	-0.15	-0.20	-0.28	-0.05	-0.20	-0.04	-0.13	-0.12
1-Butanol	-0.05	0.27	0.27	0.02	-0.11	0.02	-0.48	-0.42	-0.30	-0.16	-0.16
1-Pentanol	0.27	0.54	0.54	0.34	0.25	0.31	-0.07	-0.06	-0.20	-0.24	-0.08
1-Hexanol	0.54	0.68	0.68	0.54	0.48	0.26	0.25	0.26	0.02	-0.01	-0.01
1-Heptanol	0.68	0.54	0.54	0.45	0.41	0.23	0.54	0.40	0.14	0.01	0.06
2-Methyl-1-propanol	-0.39	-0.15	-0.17	-0.16	-0.17	-0.45	-0.37	-0.28	-0.30	-0.24	-0.29
2-Methyl-1-butanol	-0.15	-0.05	-0.22	-0.16	-0.22	-0.27	-0.07	-0.13	-0.10	-0.17	-0.31
3-Methyl-1-butanol	-0.19	-0.07	-0.07	-0.18	-0.23	-0.10	-0.37	-0.26	-0.30	-0.24	-0.21
2,2-Dimethyl-1-propanol	-0.67	-0.39	-0.39	-0.42	-0.35	-0.52	-0.47	-0.43	-0.40	-0.24	-0.29
2-Propanol	-0.47	-0.39	-0.37	-0.54	-0.46	-0.38	-0.39	-0.75	-0.66	-0.58	-0.47
2-Butanol	-0.35	-0.15	-0.15	-0.12	-0.11	-0.18	-0.07	-0.06	0.05	-0.08	-0.10
2-Pentanol	-0.07	0.15	0.15	-0.06	-0.16	-0.18	-0.35	-0.04	-0.04	-0.07	-0.15
2-Hexanol	0.15	0.25	0.25	0.14	0.09	-0.01	-0.47	-0.10	-0.09	-0.01	0.10
2-Heptanol	0.25	0.15	0.15	0.28	0.35	0.31	0.54	0.04	0.11	0.12	0.08
3-Pentanol	-0.37	-0.47	-0.68	-0.61	-0.68	-0.39	-0.19	-0.29	-0.24	-0.20	-0.23
3-Hexanol	-0.47	-0.07	-0.22	-0.17	-0.22	-0.04	0.15	0.21	0.12	0.22	0.11
2-Methyl-3-pentanol	-0.89	-1.38	-1.38	-1.04	-0.88	-0.51	-0.07	-0.11	-0.14	-0.22	-0.25
2,4-Dimethyl-3-pentanol	-1.38	-0.89	-0.89	-0.72	-0.64	-0.56	-0.89	-0.89	-0.58	-0.45	-0.40

diagonal element d_{ij} of $D(G)$ is zero. We give below the distance matrix $D(G_1)$ of the labeled hydrogen-suppressed graph G_1 of 1-butanol (Figure 1):

Figure 1. Hydrogen-suppressed graph of 1-butanol

W is calculated as:

$$W = \frac{1}{2} \sum_{i,j} d_{ij} = \sum_h g_h \quad (1)$$

where g_h is the number of unordered pairs of vertices whose distance is h .

Information-theoretic topological indices are calculated by the application of information theory on chemical graphs. An appropriate set A of n elements is derived from a molecular graph G, depending on certain structural characteristics. On the basis of an equivalence relation defined on A, the set A is partitioned into disjoint subsets A_i of order n_i ($i=1, 2, \dots, h$; $\sum n_i = n$). A probability distribution is then assigned to the set of equivalence classes:

$$\begin{aligned} A_1, A_2, \dots, A_h \\ p_1, p_2, \dots, p_h \end{aligned}$$

where $p_i = n_i/n$ is the probability that a randomly selected element of A will occur in the i^{th} subset.

The mean information content of an element of A is defined by Shannon's (1948) relation:

$$IC = - \sum_{i=1}^h p_i \log_2 p_i \quad (2)$$

The logarithm taken at base 2 measures information content in bits, and set A is then n times IC.

To account for the chemical nature of vertices and their bonding pattern, Sarkar et al. (1978) calculated information content (IC) of chemical graphs on an equivalence relation, where two atoms of the same element are considered equivalent if they possess an identical first-order topological neighborhood. Since properties of atoms, or reaction centers, are often modulated by physicochemical characteristics of distant neighbors, that is, neighbors of neighbors, it was deemed essential to extend this approach to account for higher-order neighbors of vertices. This can be accomplished by defining open spheres for all vertices of a chemical graph. One can construct such open spheres for higher integral values of r . For a particular value of r , the collection of all such open spheres $S(v,r)$, where v runs over the whole vertex set V, forms a neighborhood system of the vertices of G. A suitably defined equivalence relation can then partition V into disjoint subsets consisting of topological neighborhoods of vertices of up to r^{th} order neighbors.

This approach has been used to generate the indices of neighborhood symmetry. In this method, chemical species are symbolized by weighted linear graphs. Two vertices u_v and v_v of a molecular graph are said to be equivalent with respect to the r^{th} order neighborhood if, and only if, corresponding to each path u_0, u_1, \dots, u_r of length r , there is a distinct path v_0, v_1, \dots, v_r of the same length, such that the paths have similar edge weights, and both u_v and v_v are connected to the same number and type of atoms up to the r^{th} order bonded neighbors. The detailed equivalence relation is described in our earlier studies (Roy et al. 1984).

Once partitioning of the vertex set for a particular order of neighborhood is completed, IC_r is calculated from Equation 2. Subsequent information theoretic invariants include structural information content (SIC_r) shown in Equation 3 (Basak et al. 1980) and complementary information content (CIC_r) shown in Equation 4 (Basak and Magnuson 1983). In both equations, n is the total number of vertices of the graph:

$$SIC_r = IC_r / \log_2 n \quad (3)$$

$$CIC_r = \log_2 n - IC_r \quad (4)$$

Table 3. Topological index symbols and definitions.

I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
I_D^M	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
IC	Information content of the distance matrix partitioned by frequency of occurrences of distance h
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
O_{ORB}	Maximum order of neighborhood of vertices for I_{ORB} within the hydrogen-suppressed graph
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
IC_r	Mean information content or complexity of a graph based on the r^h ($r = 0-4$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r^h ($r = 0-4$) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r^h ($r = 0-4$) order neighborhood of vertices in a hydrogen-filled graph
$h\chi$	Path connectivity index of order $h = 0-6$
$h\chi_C$	Cluster connectivity index of order $h = 3-5$
$h\chi_{PC}$	Path-cluster connectivity index of order $h = 4-6$
$h\chi^V$	Valence path connectivity index of order $h = 0-6$
$h\chi_C^V$	Valence cluster connectivity index of order $h = 3-5$
$h\chi_{PC}^V$	Valence path-cluster connectivity index of order $h = 4-6$
P_h	Number of paths of length $h = 0-7$
J	Balaban's J index based on distance
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii

The information-theoretic index on graph distance, I_D^W is calculated from the distance matrix $D(G)$ of a chemical graph G by the method of Bonchev and Trinajstić (1977):

$$I_D^W = W \log_2 W - \sum_h g_h \cdot h \log_2 h \quad (5)$$

The mean information index, \bar{I}_D^W is found by dividing the information index I_D^W by W .

Indices developed by A.T. Balaban (1982, 1983, 1986) were calculated and used in this analysis. Balaban denoted these as J indices, which are based upon the distance sums s_i of a chemical graph. J is defined as:

$$J = q(\mu + 1)^{-1} \sum_{i,j \text{ edges}} (s_i s_j)^{-1/2} \quad (6)$$

where the cyclomatic number μ (or number of rings in the graph) is $\mu = q - n + 1$, with q adjacencies or edges and n vertices. In the original definition of J , the term s_i referred to either the row distance sum for vertex i in the distance matrix (D) or the multigraph distance matrix (M):

$$s_i = \sum_j d_{ij} \quad (7)$$

For distance matrix D , each matrix element d_{ij} represents the distance from vertex i to vertex j . The diagonal entries are all zero, and the distance between any two adjacent vertices would be one. All other entries in this matrix would be the number of edges or bonds traversed in the shortest path from i to j . To account for the periodicity of chemical properties for heteroatoms, Balaban proposed two J variants: J^X , which includes corrections for heteroatom electronegativities, and J^Y , which has corrections for heteroatom covalent radii (Balaban 1986).

Atom Pairs

Atom pairs were calculated using the method of Carhart et al (1985). An atom pair is defined as a substructure consisting of two non-hydrogen atoms i and j and their interatomic separation:

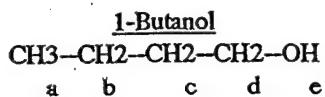
$\langle \text{atom descriptor}_i \rangle - \langle \text{separation} \rangle - \langle \text{atom descriptor}_j \rangle$

where $\langle \text{atom descriptor}_i \rangle$ contains information about the element type, number of non-hydrogen neighbors and the number of π electrons. Interatomic separation of two atoms is the number of atoms traversed in the shortest bond-by-bond path containing both atoms.

Figure 2 demonstrates the calculation of atom pairs for 1-butanol. 1-Butanol has 10 total atom pairs, 9 of which are unique. In Figure 2, X_k ($n = 1$ or 3 in this example) represents the number of non-hydrogen neighbors and the C and O are atomic symbols. These are the elements of the atom descriptors. The " $-k-$ ", $k = 2, 3, 4$, and 5 are the separation values.

The first atom pair ($CX_1 - 2 - CX_2$) corresponds to path ab , a methyl carbon with 1 non-hydrogen neighbor bonded and a methyl carbon with 2 non-hydrogen neighbors. The path length of ab is 2. Path bc and cd are identical; each consists of a path of length 2 which joins two methyl carbons, each of which has two non-hydrogen neighbors. Hence, this atom pair has a frequency of 2. Path de , involving a methyl carbon and the oxygen of the hydroxyl group, defines atom pair 3.

Figure 2. Determination of atom pairs for 1-butanol



Atom Pair	Freq. of Occurrence	Path
1. CX ₁ - 2 - CX ₂	1	ab
2. CX ₂ - 2 - CX ₂	2	bc, cd
3. CX ₂ - 2 - OX ₁	1	de
4. CX ₁ - 3 - CX ₂	1	abc
5. CX ₂ - 3 - CX ₂	1	bcd
6. CX ₂ - 3 - OX ₁	1	cde
7. CX ₁ - 4 - CX ₂	1	abcd
8. CX ₂ - 4 - OX ₁	1	bcde
9. CX ₁ - 5 - OX ₁	1	abcde

STATISTICAL ANALYSIS AND SIMILARITY MEASURES

Data Reduction

Initially, all TIs were transformed by the natural logarithm of the value of the index plus one. This was done because the scale of some TIs may be several orders of magnitude greater than others.

Principal Components Analysis (PCA)

The data analyzed in this paper may be viewed as n (number of chemicals) vectors in p (number of TIs) dimensions. The data for each set can be represented by a matrix X , which has n rows and p columns. For each of the molecular structures, the number of calculated parameters was 64 (TIs of Table 3). Each chemical is therefore represented by a point in R^64 . If each molecule is represented in R^2 , then one could plot and investigate the extent of relationship between individual parameters. In R^64 such a simple analysis is not possible. However, since many of the TIs are highly intercorrelated, the points in R^64 can likely be represented by a subspace of fewer dimensions. The method of principal components analysis (PCA), or the Karhunen-Loeve transformation, is a standard method for reduction of dimensionality (Gnanadesikan 1977). The first principal component (PC) is the line which comes closest to the points, in the sense of minimizing the sum of the squared Euclidean distances from the points to the line. The second PC is given by projections onto the basis vector orthogonal to the first PC. For points in R^p , the first r principal components give the subspace which comes closest to approximating the n points. The first PC is the first axis of the points. Successive axes are major directions orthogonal to previous axes. The PCs are the closest approximating hyperplane, and because they are calculated from Eigenvectors of a $p \times p$ matrix, the computations are relatively accessible. But there are important scaling choices, because PCs are scale dependent. To control this dependence, the most commonly used convention

is to rescale the variables so that each variable has mean zero and standard deviation one. The covariance matrix for these rescaled variables is the correlation matrix. The PCA on the TIs has been carried out using SAS software (SAS 1989).

Similarity Measures

Two measures of intermolecular similarity were used in the study of the alkanols. The methods have been described in detail previously (Basak and Grunwald 1995a) and include an associative measure using atom pairs (AP) and Euclidean distance (ED) within a five-dimensional PC space.

K-Neighbor Selection

Using the topologically-based methods described above, the intermolecular similarity of the chemicals was quantified. For each chemical, the K nearest neighbors ($K=1-5$) were then determined using the two similarity techniques. The mean observed pIC_{50} of the K-nearest neighbors for a compound was used as the estimated pIC_{50} for the compound. The correlation (r) of observed pIC_{50} with estimated pIC_{50} and the standard error (SE) of the estimates were used to assess the relative efficacy of the two similarity methods.

RESULTS

From the PCA of 64 TIs for 19 alkanols, five PCs were retained. These five PCs explain, cumulatively, 94.5% of the total variation within the TI data. Table 4 lists the Eigenvalues of the five PCs, the proportion of variance explained by each PC, the cumulative variance explained, and the two TIs most correlated with each PC. The first PC is strongly correlated with parameters which characterize the size of the molecular graph such as, P_0 and ${}^6\chi$. The second PC is highly correlated with the higher order complexity indices including CIC_4 and SIC_4 . For the third PC, the highest correlations occur with low order information theoretic complexity TIs such as IC_2 and IC_1 . The fourth PC was characterized by sixth order path-cluster and valence path-cluster connectivity indices, ${}^6\chi_{PC}$ and ${}^6\chi_{PC}$; the fifth PC, by parameters that characterize larger linear graphs, P_7 , and ${}^6\chi_{PC}$. These PC/TI correlations agree with our general expectations based on previous research (Basak and Grunwald 1994, 1995a, 1995c, 1995d). Generally, PCs and TIs correlate as follows: PC_1 with the size of the molecular graph, PC_2 with higher order complexity indices, PC_3 with cluster and path-cluster connectivity, and PC_4 with low order information theoretic indices. In this particular case, we see PC_3 and PC_4 reversed.

Table 4. Summary of the principal component analysis (PCA) using 64 TIs for 19 alkanols.

PC	Eigenvalue	% Variance	Cumulative %	1st Correlated TI	2nd Correlated TI
1	34.7	54.2	54.2	P_0	${}^6\chi$
2	16.3	25.5	79.7	CIC_4	SIC_4
3	4.9	7.7	87.4	IC_2	IC_1
4	3.5	5.5	92.9	${}^6\chi_{PC}$	${}^6\chi_{PC}$
5	1.0	1.6	94.5	P_7	${}^6\chi_{PC}$

TI: Topological indices

Table 2 presents estimates of pIC_{50} for each similarity method at each K-level (K=1-5). The atom pair method gave the best overall results. The AP standard errors fell within the range of the PC standard errors, and the correlations were all 10% to 25% higher. The best correlation for the atom pair method was 0.878 for K=1. It should be noted that results for K of 1, 2, and 3 were all very close, within 0.013 units for correlation and standard errors of within 0.01 -log units.

Table 5 reports the correlation and standard errors of pIC_{50} estimates with observed values for both the atom pair (AP) and the principal component (PC) methods. Each line of the table represents a different K level. The standard error for estimation was at its minimum of 0.17 -log units for the PC method with K=5. The correlation, however, was at its maximum of 0.878 using the AP method with K=1.

Table 5. -Log IC_{50} estimation for alkanols by K-nearest neighbors using atom pair (AP) and Euclidean distance (ED) similarity approaches

K	AP Method		ED Method	
	r	SE	r	SE
1	0.878	0.26	0.661	0.36
2	0.865	0.27	0.707	0.30
3	0.871	0.26	0.595	0.23
4	0.855	0.29	0.566	0.19
5	0.811	0.34	0.638	0.17

r: Correlation

SE: Standard error

DISCUSSION

The objective of this study was to investigate the utility of nonempirically based molecular similarity methods in estimating the inhibitory potency (pIC_{50}) of a group of aliphatic alcohols for microsomal p-hydroxylation of aniline. The result shows that the atom pair method of quantifying similarity gives a reasonable estimate of pIC_{50} values of alkanols (Table 2).

It is evident from an analysis of results in Table 5 that the AP method is superior to the ED method in predicting pIC_{50} values. This is true for K = 1-5. This indicates that atom pairs quantify structural aspects of alkanols, relevant to inhibition of aniline p-hydroxylation by microsomal cytochrome P₄₅₀, better than the Euclidean space derived from the calculated numerical graph invariants. Further work is in progress to determine the relative effectiveness of AP vis-a-vis ED methods in estimating physicochemical as well as toxicological properties of chemicals.

REFERENCES

Arcos JC (1987). Structure-activity relationships: Criteria for predicting carcinogenic activity of chemical compounds. Environ Sci Technol 21:743-745.

Auer CM, Nabholz JV, Baetcke KP (1990). Mode of action and the assessment of chemical hazards in the presence of limited data: Use of structure-activity relationships (SAR) under TSCA, Section 5. *Environ Health Perspect* 87: 183-197.

Balaban AT (1982). Highly discriminating distance-based topological index. *Chem Phys Lett* 89: 399-404.

Balaban AT (1983). Topological indices based on topological distances in molecular graphs. *Pure Appl Chem* 55: 199-206.

Balaban AT (1986). Chemical graphs. Part 48. Topological index J for heteroatom-containing molecules taking into account periodicities of element properties. *Math Chem (MATCH)* 21: 115-122.

Basak SC (1987). Use of molecular complexity indices in predictive pharmacology and toxicology: A QSAR approach. *Med Sci Res* 15: 605-609.

Basak SC, Bertelsen S, Grunwald GD (1994). Application of graph theoretical parameters in quantifying molecular similarity and structure-activity studies. *J Chem Inf Comput Sci* 34: 270-276.

Basak SC, Rosen ME, Magnuson VR (1986). Molecular topology and mutagenicity: A QSAR study of nitrosamines. *IRCS Med Sci* 14: 848-849.

Basak SC, Frane CM, Rosen ME, Magnuson VR (1987a). Molecular topology and acute toxicity: A QSAR study of monoketones. *Med Sci Res* 15: 887-888.

Basak SC, Gieschen DP, Harriss DK, Magnuson VR (1983). Physicochemical and topological correlates of enzymatic acetyl transfer reaction. *J Pharm Sci* 72: 934-937.

Basak SC, Gieschen DP, Magnuson VR (1984). A quantitative correlation of the LC_{50} values of esters in *Pimephales promelas* using physicochemical and topological parameters. *Environ Toxicol Chem* 3: 191-199.

Basak SC, Gieschen DP, Magnuson VR, Harriss DK (1982). Structure-activity relationships and pharmacokinetics: a comparative study of hydrophobicity, van der Waals' volume and topological parameters. *IRCS Med Sci* 10: 619-620.

Basak SC, Grunwald GD (1993). Use of graph invariants, volume and total surface area in predicting boiling point of alkanes. *Math Modelling Sci Computing* 2: 735-740.

Basak SC, Grunwald GD (1994). Molecular similarity and risk assessment: Analog selection and property estimation using graph invariants. *SAR and QSAR in Environ Res* 2: 289-307.

Basak SC, Grunwald GD (1995a). Estimation of lipophilicity from molecular structural similarity. *New J Chem* 19: 231-237.

Basak SC, Grunwald GD (1995b). Molecular similarity and estimation of molecular properties. *J Chem Inf Comput Sci* 35:366-372.

Basak SC, Grunwald GD (1995c). Tolerance space and molecular similarity. SAR and QSAR in Environ Res 3:265-277.

Basak SC, Grunwald GD (1995d). Predicting mutagenicity of chemicals using topological and quantum chemical parameters: A similarity based study. Chemosphere 31:2529-2546.

Basak SC, Grunwald GD (In Press a). Use of topological space and property space in selecting structural analogs. Math Modelling Sci Computing.

Basak SC, Gute BD, Grunwald GD (In Press b). Estimation of normal boiling points of haloalkanes using molecular similarity. Croat Chim Acta.

Basak SC, Harriss DK, Magnuson VR (1988a). POLLY 2.3 Copyright of the University of Minnesota.

Basak SC, Magnuson VR (1983). Molecular topology and narcosis: A quantitative structure-activity relationship (QSAR) study of alcohols using complementary information content (CIC). Arzneim Forsch/ Drug Research 33: 501-503.

Basak SC, Magnuson VR, Niemi GJ, Regal RR (1988b). Determining structural similarity of chemicals using graph-theoretic indices. Discrete Appl Math 19: 17-44.

Basak SC, Magnuson VR, Niemi GJ, Regal RR, Veith GD (1987b). Topological indices: Their nature, mutual relatedness, and applications. Mathematical Modelling 8: 300-305.

Basak SC, Niemi GJ, Veith GD (1990). Optimal characterization of structure for prediction of properties. J Math Chem 4: 185-205.

Basak SC, Niemi GJ, Veith GD (1991). Predicting properties of molecules using graph invariants. J. Math Chem 7: 243-272.

Basak SC, Roy AB, Ghosh JJ (1980). Study of the structure-function relationship of pharmacological and toxicological agents using information theory. In: Proceedings of the 2nd International Conference on mathematical modelling, II. Rolla: University of Missouri-Rolla: 851-856.

Bonchev D, Trinajstić N (1977). Information theory, distance matrix and molecular branching. J Chem Phys 67: 4517-4533.

Carhart RE, Smith DH, Venkataraghavan R (1985). Atom pairs as molecular features in structure-activity studies: Definition and applications. J Chem Inf Comput Sci 25: 64-73.

Cohen GM, Mannerling GJ (1973). Involvement of a hydrophobic site in the inhibition of the microsomal p-hydroxylation of aniline by alcohols. Mol Pharmacol 9: 383-397.

Gnanadesikan R (1977). Methods for statistical analysis of multivariate observations. New York: Wiley.

Hansch C (1976). On the structure of medicinal chemistry. J Med Chem 19: 1-6.

Johnson MA, Maggiora GM (1990). Concepts and applications of molecular similarity. New York: Wiley.

Johnson M, Basak SC, Maggiora G (1988). A characterization of molecular similarity methods for property prediction. Mathematical and Computer Modelling II: 630-635.

Kier LB, Hall LH (1986). Molecular connectivity in structure-activity analysis. Letchworth, Hertfordshire, UK: Research Studies Press.

Lajiness M (1990). Molecular similarity-based methods for selecting compounds for screening. In: Rouvray DH, ed. Computational chemical graph theory. New York: Nova Science Publishers, 299-316.

National Research Council (NRC) (1984). Toxicity testing strategies to determine needs and priorities. Washington, DC: National Academy Press.

Niemi GJ, Basak SC, Veith GD, Grunwald GD (1992). Prediction of octanol-water partition coefficient (Kow) using algorithmically-derived variables. Environ Toxicol Chem 11: 893-900.

Randić M (1975). On characterization of molecular branching. J Am Chem Soc 97: 6609-6615.

Raychaudhury C, Ray SK, Ghosh JJ, Roy AB, Basak SC (1984). Discrimination of isomeric structures using information theoretic topological indices. J Comput Chem 5: 581-588.

Roy AB, Basak SC, Harriss DK, Magnuson VR (1984). Neighborhood complexities and symmetry of chemical graphs and their biological applications. In: Xavier JR et al. eds. Mathematical modelling in science and technology. The 4th International Conference, Zurich, Switzerland, August, 1983. New York: Pergamon Press, 745-750.

Sarkar R, Roy AB, Sarkar PK (1978). Topological information content of genetic molecules-I. Math Biosci 39: 299-312.

SAS/STAT user's guide version 6, Fourth Edition, 2. (1989). Cary, North Carolina: SAS Institute Inc 846.

Shannon CE (1948). A mathematical theory of communication. Bell Sys Tech J 27: 379-423.

Toxic Substances Control Act (TSCA) (1976). Public Law 94-469, 90 Stat. 2003, October 11.

Wiener H (1947). Structural determination of paraffin boiling points. J Am Chem Soc 69: 17-20.

Wilkins CL, Randić M (1980). A graph theoretic approach to structure-property and structure-activity correlations. Theoret Chim Acta (Berl.) 58: 45-68.

ISSN-0011-1643
CCA-2366

Original Scientific Paper

Estimation of the Normal Boiling Points of Haloalkanes Using Molecular Similarity

Subhash C. Basak, Brian D. Gute, and Gregory D. Grunwald

*Natural Resources Research Institute, University of Minnesota, 5013 Miller Trunk
Highway, Duluth, MN 55811, USA*

Received April 3, 1995; revised November 3, 1995; accepted November 4, 1995

A molecular similarity measure has been used to estimate the normal boiling points of a set of 267 haloalkanes with 1-4 carbon atoms. Molecular similarity/dissimilarity was quantified in terms of Euclidean distances of molecules in the eight dimensional principal component space derived from fifty-nine topological indices. Correlation coefficients between the experimental and estimated boiling points ranged from 0.854 to 0.943 in the K -nearest neighbor estimation of boiling points using a different number of nearest neighbors ($K = 1-10, 15, 20, 25$).

INTRODUCTION

The use of structural analogy as a tool to classify chemicals, as well as predict the behaviour of chemical species, is as old as chemistry. In 1819, Mitscherlich¹ described the phenomenon of isomorphism, in which substitution of one atom by another leads to similar lattice structures. At the turn of this century, Langmuir² observed that isosteric chemical species, those which contain the same total number of atoms and electrons, have very similar properties. Members of isosteric pairs, like N₂-CO and N₂O-CO₂, have many similar physical constants.³ The structural similarity of the isosteric amino acids valine and threonine poses some interesting problems in the protein synthesis mechanism of cells. Being sterically similar, valine and threonine may be charged to the same tRNA. The incorrectly formed aminoacyl adenylate and aminoacyl tRNA are discriminated and destroyed *via* a »double sieve«, involving steric exclusion and ineffective binding, before they are used in protein synthesis.⁴

Similarity plays an important role in biological activity. The enzyme dihydrofolate reductase normally facilitates the reduction of dihydrofolate to tetrahydrofolate. Methotrexate, a compound whose structure is similar to dihydrofolate, inhibits the action of the reductase.⁵ Competitive inhibition of enzymes can also result from interaction of the enzyme with transition state analogs of the substrate. For example, proline racemase from *Clostridium sticklandii* preferentially binds the transition state of proline. As a result, the racemase is subject to inhibition by compounds which are structural analogs of the transition state of proline, such as pyrrole-2-carboxylate and pyrrolidine-2-carboxylate, which bind to the enzyme with a much greater affinity than does proline.⁶ Furthermore, the structural similarity between a macromolecular biotarget and its antiidiotypic antibody is believed to be the reason for the use of such antibodies as model receptors in the screening of chemicals for drug discovery.⁷

The last decade has seen an upsurge of interest in the development of similarity measures and their applications in chemical research, drug design, and toxicology.⁸⁻²⁵ Such methods are based on different representations of chemical species, viz., topological, geometrical, quantum chemical, etc. In drug design, similarity searching of databases is used to identify potential leads. Also, dissimilarity based methods are used to select chemicals for screening in the drug discovery process.¹¹ In toxicology, structural and functional analogy are used to assess the ecological and human health risk of the new and existing chemicals.²⁶⁻²⁸

In the United States, the majority of chemicals submitted to the Environmental Protection Agency (USEPA) for registration do not have any test data.²⁷ One of the methods used by regulators for the hazard assessment of such chemicals is to select their analogs and, subsequently, estimate the hazard of the chemical of interest from the hazard of the analogs. Such selection of analogs is often done subjectively by individual experts on the basis of an intuitive notion of similarity.²⁷ In USEPA's approach to ecological risk assessment, class specific QSARs are preferred over the use of analogs, although in human health hazard assessment, analog-based estimation of toxic potential is still the most important factor.²⁸

Rapid selection of analogs for drug design and hazard assessment requires automated methods that are computationally feasible. Similarity methods based on parameters that can be calculated directly from molecular structure fall into this category.⁸⁻²⁵ Topological indices derived from a molecular graph comprise a set of parameters which can be computed for any chemical structure.²⁹

In some of our recent studies, we have developed novel methods of quantifying molecular similarity using topological indices and substructural features like atom pairs.¹³⁻²² We have applied similarity techniques in the selection of analogs and in the estimation of molecular properties such as

boiling point, lipophilicity, and mutagenicity for different sets of chemicals. In this paper, we have carried out a similarity based estimation of the normal boiling point for a set of 267 chlorofluorocarbons (CFCs) using a similarity method based on topological indices.

DATABASES

The data analyzed in this study consist of the normal boiling points for 267 CFCs with 1–4 carbon atoms. These data were originally collected from Beilstein's *Handbuch der Organischen Chemie*, the *CRC Handbook of Chemistry and Physics*, Heilbron's *Dictionary of Organic Compounds* and Smith and Srivastava's *Thermodynamic Data for Pure Compounds, Part B* for use in several studies by Balaban *et al.*^{30,31} For our purposes, the subset of 276 CFC's³⁰ was further reduced to 267, to remove outliers. Nine compounds whose normal boiling points were more than two standard deviations from the mean boiling point of the group were removed. This was done to enhance the estimation by removing compounds that had only one or two neighbours which would give reasonable estimates of boiling point. Table I is a listing of the compounds used in this study and their normal boiling points.

METHODS

Calculation of Topological Indices

The fifty-nine topological indices used in this study were calculated using POLLY 2.3 which uses the SMILES line notation input of chemical structures.³² The TIs calculated are listed in Table II and include the Wiener index calculated by the method of Wiener,³³ connectivity indices as calculated by Randić³⁴ and by Kier and Hall,³⁵ information theoretic indices defined on distance matrices of graphs using the methods of Bonchev and Trinajstić³⁶ as well as those of Raychaudhury *et al.*,³⁷ parameters derived on the neighbourhood complexity of vertices in hydrogen-filled molecular graphs,^{38–41} path lengths, and Balaban's *J* indices.^{42–44}

Data Reduction

Initially, all TIs were transformed by the natural log of the TI plus one. The natural logarithm transformation was done because some TIs may be several orders of magnitude greater than others. One was added before the log transformation since many of the TIs may be zero. Principal component analysis (PCA) was used to reduce the dimensionality of the set of 59 topological indices (TIs). With PCA, linear combinations of the TIs, called prin-

TABLE I

Normal boiling points of 267 haloalkanes with 1-4 carbon atoms

No.	Chemical Name	Normal Boiling Point	Est. Boiling Point	Residual Boiling Point
1	carbon tetrachloride	76.7	33.3	43.4
2	trichloromethane	61.2	28.6	32.6
3	dichloromethane	39.8	7.1	32.7
4	trichlorofluoromethane	23.7	1.3	22.4
5	dichlorofluoromethane	8.9	4.3	4.6
6	chlorofluoromethane	-9.1	3.5	-12.6
7	chloromethane	-24.3	-9.3	-15.0
8	dichlorodifluoromethane	-29.8	6.9	-36.7
9	chlorodifluoromethane	-40.8	-8.4	-32.4
10	difluoromethane	-51.7	12.0	-63.7
11	hexachloroethane	184.4	146.6	37.8
12	1,1,1,2,2-pentachloro-2-fluoroethane	137.9	136.2	1.7
13	1,1,1,2-tetrachloro-2-fluoroethane	117.0	106.9	10.1
14	1,1,2,2-tetrachloro-1-fluoroethane	116.6	97.7	18.9
15	1,1,2-trichloroethane	113.7	78.7	35.0
16	1,1,2-trichloro-2-fluoroethane	102.4	68.8	33.6
17	1,1,2,2-tetrachloro-1,2-difluoroethane	92.7	55.0	37.7
18	1,1,1-trichloroethane	74.0	25.2	48.8
19	1,2-dichloro-1-fluoroethane	73.8	71.4	2.4
20	1,1,2-trichloro-1,2-difluoroethane	72.5	60.4	12.1
21	1,2-dichloro-1,2-difluoroethane	58.5	33.9	24.6
22	1,1-dichloroethane	57.2	4.8	52.4
23	1,1,1-trichloro-2,2,2-trifluoroethane	45.8	57.5	-11.7
24	1,2-dichloro-1,1-difluoroethane	46.6	43.8	2.8
25	2-chloro-1,1-difluoroethane	35.1	57.6	-22.5
26	1,1-dichloro-1-fluoroethane	32.0	15.2	16.8
27	2,2-dichloro-1,1,1-trifluoroethane	28.7	42.3	-13.6
28	1-chloro-1-fluoroethane	16.1	32.6	-16.5
29	chloroethane	12.3	7.6	4.7
30	1-chloro-1,1,2-trifluoroethane	12.0	20.0	-8.0
31	2-chloro-1,1,1-trifluoroethane	6.9	21.1	-14.2
32	1,1,2-trifluoroethane	5.0	27.7	-22.7
33	1,2-dichloro-1,1,2,2-tetrafluoroethane	3.6	40.6	-37.0
34	2,2-dichloro-1,1,1,2-tetrafluoroethane	3.6	29.1	-25.5
35	1-chloro-1,1,2,2-tetrafluoroethane	-12.0	24.1	-36.1
36	1,1,2,2-tetrafluoroethane	-22.8	63.6	-86.4
37	1,1-difluoroethane	-25.8	27.8	-53.6
38	1,1,1,2-tetrafluoroethane	-26.1	-1.1	-25.0
39	fluoroethane	-37.8	17.6	-55.4
40	1,1,1-trifluoroethane	-47.3	-8.5	-38.8
41	1,1,1,2,2-pentafluoroethane	-48.3	-14.4	-33.9
42	1,1,2,2,3,3-hexachloropropane	218.5	201.4	17.1
43	1,1,1,2,2,3-hexachloropropane	218.0	199.9	18.1
44	1,1,1,2,3,3-hexachloro-2,3-difluoropropane	196.0	183.8	12.2
45	1,1,1,2,2,3-hexachloro-3,3-difluoropropane	193.4	197.4	-4.0

TABLE I
(continuing)

No.	Chemical Name	Normal Boiling Point	Est. Boiling Point	Residual Boiling Point
46	1,1,1,2,2-pentachloro-3,3-difluoropropane	175.0	178.0	-3.0
47	1,1,1,3,3-pentachloro-2,2-difluoropropane	174.0	147.8	26.2
48	1,1,2,2-tetrachloropropane	165.5	173.1	-7.6
49	1,1,3,3-tetrachloropropane	161.9	170.6	-8.7
50	1,2,3-trichloropropane	156.8	153.4	3.4
51	1,1,2,3,3-pentachloro-1,2,3-trifluoropropane	154.7	128.6	26.1
52	1,1,2,2-tetrachloropropane	153.0	151.6	1.4
53	1,1,2,2,3-pentachloro-1,3,3-trifluoropropane	152.3	156.8	-4.5
54	1,1,1,3-tetrachloro-2,2-difluoropropane	151.2	130.7	20.5
55	1,1,1,2-tetrachloropropane	150.4	152.1	-1.7
56	1,1,2,2-tetrachloro-3,3-difluoropropane	147.6	132.1	15.5
57	1,1,3-trichloropropane	145.5	140.9	4.6
58	1,1,2,2-tetrachloro-1,3,3-trifluoropropane	134.6	126.9	7.7
59	1,2,3-trichloro-2-fluoropropane	130.8	111.9	18.9
60	1,1,2,3-tetrachloro-2,3,3-trifluoropropane	129.8	106.8	23.0
61	1,1,3-trichloro-2,2-difluoropropane	127.3	99.2	28.1
62	1,1,2,2-tetrachloro-3,3,3-trifluoropropane	126.2	132.3	-6.1
63	1,1,2-trichloro-2-fluoropropane	116.7	99.2	17.5
64	1,1,3,3-tetrachloro-1,2,2,3-tetrafluoropropane	114.0	105.0	9.0
65	1,1,1,3-tetrachloro-2,2,3,3-tetrafluoropropane	113.9	104.3	9.6
66	1,1,2-trichloro-1-fluoropropane	113.5	99.9	13.6
67	1,1,1,2-tetrachloro-2,3,3,3-tetrafluoropropane	112.5	121.2	-8.7
68	1,1,2,2-tetrachloro-1,3,3,3-tetrafluoropropane	112.3	125.2	-12.9
69	1,2,2,3-tetrachloro-1,1,3,3-tetrafluoropropane	112.2	113.4	-1.2
70	1,1,3-trichloro-1,2,2-trifluoropropane	109.5	87.2	22.3
71	1,1,1-trichloropropane	108.0	105.9	2.1
72	1,2,2-trichloro-3,3,3-trifluoropropane	104.5	120.8	-16.3
73	1,3-dichloro-2,2-difluoropropane	96.7	93.6	3.1
74	1,3,3-trichloro-1,1,2,2-tetrafluoropropane	91.8	90.2	1.6
75	1,2,2-trichloro-1,1-difluoropropane	90.2	86.5	3.7
76	1,2,3-trichloro-1,1,2,3-tetrafluoropropane	90.0	84.5	5.5
77	2,3-dichloro-1,1,2,3-tetrafluoropropane	89.8	72.2	17.6
78	1,2,3-trichloro-1,1,3,3-tetrafluoropropane	88.0	76.7	11.3
79	1-chloro-3-fluoropropane	81.0	101.7	-20.7
80	1,2,3-trichloro-1,1,2,3,3-pentafluoropropane	73.7	65.4	8.3
81	2,3,3-trichloro-1,1,1,2,3-pentafluoropropane	73.4	65.4	8.0
82	1,3,3-trichloro-1,1,2,2,3-pentafluoropropane	73.0	65.5	7.5
83	2,2,3-trichloro-1,1,1,3,3-pentafluoropropane	72.0	80.6	-8.6
84	1,2-dichloro-1,1-difluoropropane	70.0	82.1	-12.1
85	2,2-dichloropropane	69.3	34.3	35.0
86	1-chloro-2-fluoropropane	68.5	70.7	-2.2
87	1,1-dichloro-1-fluoropropane	66.6	77.6	-11.0
88	1-chloro-2,2-difluoropropane	55.1	44.0	11.1
89	1-chloro-1,2-difluoropropane	52.9	72.8	-19.9
90	2,2-dichloro-1,1,1-trifluoropropane	48.8	72.0	-23.2

TABLE I
(continuing)

No.	Chemical Name	Normal Boiling Point	Est. Boiling Point	Residual Boiling Point
91	1-chloropropane	46.6	43.8	2.8
92	3,3-dichloro-1,1,2,2-pentafluoropropane	45.5	58.8	-13.3
93	1,3-difluoropropane	41.6	88.8	-47.2
94	2-chloropropane	35.7	31.3	4.4
95	1,3-dichloro-1,1,2,2,3,3-hexafluoropropane	35.7	33.8	1.9
96	2-chloro-2-fluoropropane	35.2	25.5	9.7
97	3,3-dichloro-1,1,1,2,2,3-hexafluoropropane	35.0	50.5	-15.5
98	3-chloro-1,1,1,2,2-pentafluoropropane	27.6	42.1	-14.5
99	1-chloro-1,1-difluoropropane	25.4	50.0	-24.6
100	1-chloro-1,1,2,2,3,3-hexafluoropropane	21.0	32.2	-11.2
101	1,1,1,2,3-pentafluoropropane	20.0	20.6	-0.6
102	1,1,2,2,3,3-hexafluoropropane	10.5	11.2	-0.7
103	1,1-difluoropropane	7.5	61.1	-53.6
104	1,1,1,2,3,3-hexafluoropropane	5.0	9.0	-4.0
105	1,1,1,2,2,3-hexafluoropropane	1.2	5.3	-4.1
106	2-chloro-1,1,1,2,3,3-heptafluoropropane	-2.0	12.6	-14.6
107	2-fluoropropane	-9.7	18.1	-27.8
108	1,1,1-trifluoropropane	-12.5	7.1	-19.6
109	1,1,1,2,3,3-heptafluoropropane	-19.0	20.6	-39.6
110	1-chloro-2-fluoroethane	53.0	48.7	4.3
111	1-chloro-1,1-difluoroethane	-9.8	-1.9	-7.9
112	1-chloro-1,1,2,2-pentafluoroethane	-38.0	4.8	-42.8
113	1,2-dichloroethane	83.5	50.5	33.0
114	1,1-dichloro-2,2-difluoroethane	60.0	57.1	2.9
115	1,1-dichloro-1,2,2-trifluoroethane	30.2	41.3	-11.1
116	1,2-dichloro-1,1,2-trifluoroethane	28.2	41.7	-13.5
117	1,1,2-trichloro-1-fluoroethane	88.5	56.5	32.0
118	1,1,1-trichloro-2,2-difluoroethane	73.0	68.1	4.9
119	1,1,2-trichloro-2,2-difluoroethane	71.2	62.3	8.9
120	1,1,2-trichloro-1,2,2-trifluoroethane	47.6	47.4	0.2
121	1,1,1,2-tetrachloroethane	130.5	98.3	32.2
122	1,1,2,2-tetrachloroethane	146.3	61.7	84.6
123	1,1,1,2-tetrachloro-2,2-difluoroethane	91.6	102.0	-10.4
124	1,1,1,2,2-pentachloroethane	161.9	149.5	12.4
125	1,1,2,2,3-pentachloropropane	196.0	193.0	3.0
126	1,1,2,3,3-pentachloropropane	199.0	184.4	14.6
127	1,1,2,2,3-pentachloro-3,3-difluoropropane	168.4	148.0	20.4
128	1,1,2,3,3-pentachloro-1,3-difluoropropane	167.4	130.5	36.9
129	1,1,1,2,2-pentachloro-3,3,3-trifluoropropane	153.0	172.7	-19.7
130	1,1,1,2,3-pentachloro-2,3,3-trifluoropropane	153.3	145.2	8.1
131	1,1,1,3,3-pentachloro-2,2,3-trifluoropropane	153.0	137.3	15.7
132	1,1,1,2,3,3-hexachloropropane	217.0	207.2	9.8
133	1,1,1,3,3,3-hexachloropropane	206.0	151.7	54.3
134	1,1,1,2,2,3-hexachloro-3-fluoropropane	210.0	201.5	8.5
135	1,1,1,2,3,3-hexachloro-3-fluoropropane	207.0	178.7	28.3

TABLE I

(continuing)

No.	Chemical Name	Normal Boiling Point	Est. Boiling Point	Residual Boiling Point
136	1,1,2,2,3,3-hexachloro-1-fluoropropane	210.0	185.3	24.7
137	1,1,1,3,3,3-hexachloro-2,2-difluoropropane	194.2	148.5	45.7
138	1,1,2,2,3,3-hexachloro-1,3-difluoropropane	194.2	181.4	12.8
139	1,2-dichloro-1,1,2,3,3-pentafluoropropane	56.3	65.1	-8.8
140	2,3-dichloro-1,1,1,2,3-pentafluoropropane	56.0	65.6	-9.6
141	1,1,2-trichloropropane	133.0	94.7	38.3
142	1,2,2-trichloropropane	122.0	103.1	18.9
143	1,1,1-trichloro-2,2-difluoropropane	102.0	75.0	27.0
144	1,2,2-trichloro-1,1,3,3-tetrafluoropropane	92.0	99.4	-7.4
145	3,3,3-trichloro-1,1,1,2,2-pentafluoropropane	70.5	89.4	-18.9
146	1,1,2-trichloro-1,2-difluoropropane	97.7	74.0	23.7
147	1,1,3-trichloro-3,3-difluoropropane	107.8	70.3	37.5
148	3-chloro-1,1,1,3,3-pentafluoropropane	28.4	63.7	-35.3
149	2-chloro-1,1,1,3,3-hexafluoropropane	15.5	20.3	-4.8
150	3-chloro-1,1,1,2,2,3-heptafluoropropane	-2.5	15.0	-17.5
151	3-chloro-1,1,1,2,2,3-hexafluoropropane	20.0	22.2	-2.2
152	1,1-dichloropropane	88.1	91.8	-3.7
153	1,2-dichloropropane	96.0	90.3	5.7
154	1,3-dichloropropane	120.8	105.4	15.4
155	1,2-dichloro-2-fluoropropane	88.6	57.5	31.1
156	1,2-dichloro-1-fluoropropane	93.0	64.8	28.2
157	1,1-dichloro-2,2-difluoropropane	79.0	80.3	-1.3
158	1,3-dichloro-1,1-difluoropropane	80.8	79.6	1.2
159	1,1-dichloro-1,2,2-trifluoropropane	60.2	62.4	-2.2
160	3,3-dichloro-1,1,1-trifluoropropane	72.4	81.3	-8.9
161	1,2-dichloro-1,1,2-trifluoropropane	55.6	63.1	-7.5
162	2,3-dichloro-1,1,1-trifluoropropane	76.7	99.6	-22.9
163	1,3-dichloro-1,1,2,2-tetrafluoropropane	68.2	71.8	-3.6
164	2,3-dichloro-1,1,1,3,3-pentafluoropropane	50.4	70.3	-20.0
165	2,3-dichloro-1,1,1,2,3,3-hexafluoropropane	34.7	40.0	-5.3
166	1,2,3-trichloro-1,1-difluoropropane	114.3	102.0	12.3
167	1,1,1-trichloro-3,3,3-trifluoropropane	95.1	109.6	-14.5
168	1,1,2-trichloro-3,3,3-trifluoropropane	106.8	103.0	3.8
169	2,3,3-trichloro-1,1,1,3-tetrafluoropropane	87.2	91.4	-4.2
170	1,1,3-trichloro-1,2,2,3-tetrafluoropropane	90.5	109.6	-19.1
171	1,1,1,3-tetrachloropropane	158.0	142.6	15.4
172	1,1,2,3-tetrachloropropane	180.0	156.6	23.4
173	1,1,1,2-tetrachloro-2-fluoropropane	139.6	113.5	26.1
174	1,1,2,2-tetrachloro-1-fluoropropane	135.0	114.4	20.6
175	1,1,1,3-tetrachloro-3,3-difluoropropane	132.0	112.1	19.9
176	1,1,1,2-tetrachloro-3,3,3-trifluoropropane	125.1	141.4	-16.3
177	1,1,2,3-tetrachloro-1,3,3-trifluoropropane	128.7	111.9	16.8
178	1,1,3,3-tetrachloro-2,2,3-trifluoropropane	127.0	105.6	21.4
179	1,1,2,3-tetrachloro-1,2,3,3-tetrafluoropropane	112.5	97.0	15.5
180	1-fluoropropane	-2.3	24.7	-27.0

TABLE I
(continuing)

No.	Chemical Name	Normal Boiling Point	Est. Boiling Point	Residual Boiling Point
181	octafluoropropane	-38.0	-0.5	-37.5
182	2,2-difluoropropane	-0.5	11.3	-11.8
183	1,1,1,3-tetrafluoropropane	29.4	23.3	6.1
184	1,1,1,3,3,3-hexafluoropropane	0.8	18.2	-17.4
185	1,1,1,2,2,3,3-heptafluoropropane	-17.0	11.7	-28.7
186	1-chloro-1-fluoropropane	48.0	74.8	-26.8
187	3-chloro-1,1,1-trifluoropropane	45.1	65.1	-19.0
188	2-chloro-1,1-difluoropropane	52.0	74.7	-22.7
189	2-chloro-1,1,1-trifluoropropane	30.0	47.7	-17.7
190	1-fluorobutane	32.2	66.7	-34.5
191	2-fluorobutane	24.7	46.3	-21.6
192	1,1,1,2,4,4,4-octafluorobutane	18.0	59.5	-41.5
193	1,1,2,2,3,3,4,4-octafluorobutane	43.0	38.1	4.9
194	1,1,1,2,2,3,3,4,4-nonfluorobutane	14.0	55.4	-41.4
195	decafluorobutane	-2.0	43.0	-45.0
196	1-chlorobutane	78.5	90.1	-11.6
197	2-chlorobutane	68.5	58.6	9.9
198	1-chloro-4-fluorobutane	115.0	109.7	5.3
199	1-chloro-1,1-difluorobutane	55.5	56.1	-0.6
200	3-chloro-1,1,1-trifluorobutane	66.0	72.3	-6.3
201	1-chloro-1,1,3,3-tetrafluorobutane	70.5	78.0	-7.5
202	2-chloro-1,1,1,3,3-hexafluorobutane	51.0	74.2	-23.2
203	4-chloro-1,1,1,2,2,3,3-heptafluorobutane	54.0	49.3	4.7
204	4-chloro-1,1,1,2,2,3,3,4,4-nonafluorobutane	30.0	45.0	-15.0
205	1,1-dichlorobutane	115.0	145.0	-30.0
206	1,2-dichlorobutane	123.5	150.2	-26.7
207	1,3-dichlorobutane	133.0	140.8	-7.8
208	1,4-dichlorobutane	155.0	130.8	24.2
209	1,3-dichloro-1,1,3-trifluorobutane	129.0	80.9	48.1
210	3,4-dichloro-1,1,1,2,2,3-hexafluorobutane	72.0	78.5	-6.5
211	1,4-dichloro-1,1,3-trifluorobutane	118.5	97.5	21.0
212	2,3-dichloro-1,1,1,4,4,4-hexafluorobutane	78.0	71.9	6.1
213	4,4-dichloro-1,1,1,2,2,3,3-heptafluorobutane	76.5	72.0	4.5
214	4,4-dichloro-1,1,1,2,2,3,3-octafluorobutane	62.8	71.1	-8.3
215	3,4-dichloro-1,1,1,2,2,3,4,4-octafluorobutane	66.0	70.6	-4.6
216	1,4-dichloro-1,1,2,2,3,3,4,4-octafluorobutane	64.0	71.3	-7.3
217	2,2-dichloro-1,1,1,3,3,4,4-octafluorobutane	64.0	70.5	-6.5
218	2,3-dichloro-1,1,1,2,3,4,4-octafluorobutane	64.0	76.3	-12.3
219	1,1,1-trichlorobutane	133.5	137.9	-4.4
220	1,1,2-trichlorobutane	156.8	143.5	13.3
221	1,1,3-trichlorobutane	153.8	143.3	10.5
222	1,1,4-trichlorobutane	183.8	133.3	50.5
223	2,2,3-trichloro-1,1,1,4,4,4-hexafluorobutane	104.0	107.7	-3.7
224	4,4,4-trichloro-1,1,1,2,2,3,3-heptafluorobutane	96.5	85.4	11.1
225	1,3,4-trichloro-1,1,1,2,2,3,4,4-heptafluorobutane	99.0	91.3	7.7

TABLE I
(continuing)

No.	Chemical Name	Normal Boiling Point	Est. Boiling Point	Residual Boiling Point
226	2,2,3-trichloro-1,1,1,3,4,4,4-heptafluorobutane	97.4	77.7	19.7
227	1,1,4,4-tetrachlorobutane	200.0	148.7	51.3
228	1,2,4,4-tetrachloro-1,1,2,3,3,4-hexafluorobutane	134.0	92.6	41.4
229	1,2,3,4-tetrachloro-1,1,2,3,4,4-hexafluorobutane	134.0	85.2	48.8
230	1,1,2,3,4,4-hexachloro-1,2,3,4-tetrafluorobutane	208.0	113.2	94.8
231	1-chloroisobutane	68.3	60.5	7.8
232	2-chloroisobutane	50.7	38.5	12.2
233	1-chloro-1-fluoroisobutane	82.5	109.8	-27.3
234	1,1-dichloroisobutane	105.0	107.4	-2.4
235	1,2-dichloroisobutane	106.5	99.1	7.4
236	1,3-dichloroisobutane	136.0	134.6	1.4
237	1,1-dichloro-1-fluoroisobutane	107.0	116.1	-9.1
238	1,2,3-trichloroisobutane	163.0	146.0	17.0
239	1,1,2,3-tetrachloroisobutane	191.0	185.2	5.8
240	1,2,3-trichloro-2-chloromethylpropane	211.0	183.3	27.7
241	1,1,2,3-tetrachloro-2-chloromethylpropane	227.0	204.3	22.7
242	1-fluoroisobutane	16.0	56.1	-40.1
243	2-fluoroisobutane	12.0	38.1	-26.1
244	1,1,1,3,3,3-hexafluoroisobutane	21.5	25.5	-4.0
245	1,1,1,3,3,3-hexafluoro-2-fluoromethylpropane	40.0	18.9	21.1
246	1,1,1,3,3,3-hexafluoro-2-difluoromethylpropane	33.0	20.3	12.7
247	1,1,1,3,3,3-hexafluoro-2-trifluoromethylpropane	12.0	6.0	6.0
248	decafluoroisobutane	-0.3	3.6	-3.9
249	3-chloro-1,1,1,3,3-pentafluoroisobutane	59.0	68.6	-9.6
250	1,1,1,3,3,3-hexafluoro-2-chloromethylpropane	58.0	39.6	18.4
251	2,3-dichloro-1,1,1-trifluoroisobutane	93.5	101.0	-7.5
252	2,3-dichloro-1,1,1,3,3-pentafluoroisobutane	75.3	91.2	-15.9
253	2,3-dichloro-1,1,1,3,3-pentafluoro-2-trifluoromethylpropane	65.0	65.8	-0.8
254	1,1,2-trichloroisobutane	163.0	143.1	19.9
255	1,2,3-trichloro-1,1-difluoroisobutane	132.0	114.7	17.3
256	2,3,3-trichloro-1,1,1-trifluoroisobutane	123.7	124.6	-0.9
267	1,1,1,3,3,3-hexafluoro-2-trichloromethylpropane	107.0	106.7	0.3
258	1,1,1,2-tetrachloro-3,3,3-trifluoroisobutane	148.5	148.7	-0.2
259	1,1,1,2,3-pentachloroisobutane	211.0	201.3	9.7
260	1-chloro-1,1,2,2-tetrafluoropropane	19.9	27.5	-7.6
261	1,1,1-trichloropropane	104.0	99.6	4.4
262	2,3-dichlorobutane	116.0	105.2	10.8
263	2,2,3-trichlorobutane	143.0	152.3	-9.3
264	1,2,3-trichlorobutane	166.0	141.7	24.3
265	1,4-difluorobutane	77.8	122.0	-44.2
266	2,2-difluorobutane	30.9	40.0	-9.1
267	1,2-difluoroethane	26.0	40.7	-14.7

TABLE II
Topological index symbols and definitions

I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
\bar{I}_D^W	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
\bar{I}^C	Information content of the distance matrix partitioned by the frequency of occurrences of distance h
O	Order of neighbourhood when IC_r reaches its maximum value for the hydrogen-filled graph
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighbourhood of vertices
O_{ORB}	Maximum order of neighbourhood of vertices for I_{ORB} within the hydrogen-suppressed graph
M_1	A Zagreb group parameter = sum of the square of degree over all vertices
M_2	A Zagreb group parameter = sum of the cross-product of degrees over all neighbouring (connected) vertices
IC_r	Mean information content or complexity of a graph based on the r^{th} ($r = 0-3$) order neighbourhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for the r^{th} ($r = 0-3$) order neighbourhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for the r^{th} ($r = 0-3$) order neighbourhood of vertices in a hydrogen-filled graph
$h\chi$	Path connectivity index of the order $h = 0-5$
$h\chi_C$	Cluster connectivity index of the order $h = 3-6$
$h\chi_{PC}$	Path-cluster connectivity index of the order $h = 4-6$
$h\chi^V$	Valence path connectivity index of the order $h = 0-5$
$h\chi_C^V$	Valence cluster connectivity index of the order $h = 3-6$
$h\chi_{PC}^V$	Valence path-cluster connectivity index of the order $h = 4-6$
P_h	Number of paths of length $h = 0-5$
J	Balaban's J index based on distance
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii

incipal components (PCs) are derived from the correlation matrix. The first PC has the largest variance, or eigenvalue, of the linear combination of TIs. Each subsequent PC explains the maximal index variance orthogonal to previous PCs. With 59 TIs available, 59 PCs can be generated. For this study, PCs with an eigenvalue greater than one were retained. The PCA analysis and selection of PCs was accomplished using the SAS procedure PRINCOMP.⁴⁵ Basak *et al.*¹³ provide more detail on this approach.

Computation of Similarity

Intermolecular similarity was measured by the Euclidean distance (ED) within an *n*-dimensional space. This *n*-dimensional space consisted of orthogonal variables (PCs) derived from the TIs. ED between the molecule's *i* and *j* is defined as:

$$ED_{ij} = \left[\sum_{k=1}^n (D_{ik} - D_{jk})^2 \right]^{1/2}$$

where *n* equals the number of dimensions retained from PCA. D_{ik} and D_{jk} are the data values of the *k*th dimension for chemicals *i* and *j*, respectively.

K-nearest Neighbour Selection and Boiling Point Estimation

Following the quantification of the intermolecular similarity of the CFCs, the *K*-nearest neighbours (*K* = 1–10, 15, 20, 25) were determined on the basis of ED. The mean observed boiling point of the *K*-nearest neighbours for a compound was used as the estimated boiling point and the standard error (s.e.) of the estimates were used to assess the efficacy of this similarity method.

RESULTS

From the PCA of 59 TIs for 267 CFCs, eight PCs with eigenvalues greater than one were retained. These eight PCs explained, cumulatively, 95.0% of the total variation within the TI data. Table III lists the eigenvalues of the eight PCs, the proportion of variance explained by each PC, and the cumulative variance explained. In addition, Table III lists the two TIs most correlated with each PC. The first PC is strongly correlated with the parameters that characterize the size of the molecular graphs and the increasing number of chlorofluoro substitutions, *viz.* P_0 (number of atoms) and P_1 (number of bonds). The second PC is highly correlated with higher order complexity indices including SIC_2 and CIC_2 . For the third PC, the highest

TABLE III

Summary of the principal components of 59 TIs for the 267 haloalkanes and the correlation coefficients of the two most correlated with each principal component

PC	Eigenvalue	Percent of variance	Cumulative percent	First correlated TI	Second correlated TI
1	31.9	54.0	54.0	P_0	0.982
2	8.7	14.8	68.8	SIC_2	0.949
3	5.2	8.8	77.6	$^4\chi_C^v$	-0.668
4	3.6	6.1	83.7	$^1\chi^v$	0.475
5	2.1	3.6	87.3	$^1\chi^v$	0.495
6	1.9	3.2	90.5	P_5	0.579
7	1.5	2.5	93.0	$^2\chi^v$	0.282
8	1.2	2.0	95.0	$^4\chi_C^v$	0.324
				H^v	-0.313

correlations occur with the valence cluster connectivity TIs such as $^4\chi_C^v$ and $^3\chi_C^v$. The fourth PC was characterized by lower order valence path connectivity indices such as $^1\chi^v$ and $^3\chi^v$ and the fifth PC by the higher order valence path connectivity indices such as $^5\chi^v$ and $^4\chi^v$. Interpretation beyond the fifth level PC becomes more difficult, as it can be seen in Table III. These PC/TI correlations agree with our expectations based on previous research.^{16,17,19,20} Generally, PCs and TIs correlate as follows: PC_1 with the size of the molecular graph, PC_2 with higher order complexity indices, PC_3 with cluster and path-cluster connectivity, and PC_4 with low order information theoretic indices.

TABLE IV

Summary of the K -nearest neighbour normal boiling point estimation for 267 chlorofluoro-hydrocarbons

K	r	s.e. (°C)
1	0.854	33.2
2	0.908	26.4
3	0.923	24.5
4	0.927	24.2
5	0.933	23.7
6	0.934	24.3
7	0.934	24.3
8	0.936	24.3
9	0.939	24.4
10	0.939	24.7
15	0.936	26.2
20	0.936	27.7
25	0.943	28.0

Table IV reports the correlation and standard errors of boiling point estimates obtained by the K -nearest neighbour estimation with the observed boiling point values. Each line of the table represents a different K level. The standard error for estimation was at its minimum of 23.7 °C for $K = 5$. The correlation, however, continued an upward trend as K increased.

DISCUSSION

The goal of this paper was to investigate the usefulness of general similarity methods based on graph invariants in the prediction of the boiling points of a set of 267 chlorofluorocarbons. To this end, we used Euclidean distance in an eight dimensional PC-space as the measure of structural similarity/dissimilarity of CFCs. The results in Table IV show that the best estimates of the normal boiling point are obtained at $K = 5$. Our previous studies on similarity-based prediction of properties like lipophilicity,¹⁷ boiling point,^{16, 19} and mutagenicity^{16, 19, 20} have shown that a small number of neighbours ($K = 5-10$) will usually give the best results in property estimation.

Comparison of the K -nearest neighbour estimates reported in this paper with previous studies on the same set of CFCs shows that similarity-based estimates are inferior to predictions derived by neural net models.³⁰ In the neural net model, parametrization was done with an eye to specific structural features of CFCs. In contrast, the PC-based similarity approach used a set of general structural parameters which quantify such structural features of chemical graphs as size, shape, degree of branching, *etc.* Yet, similarity methods based on such graph theoretic parameters give a reasonably good estimate of the normal boiling point of CFCs analyzed in this paper. The usefulness of the similarity approach depends on the context, *i.e.* what level of accuracy is required.

In risk assessment, molecular similarity is used in the selection of analogs of chemicals for hazard estimation. Very often, one has to do rapid estimation of a large number of properties. Such estimations should be based on parameters that can be algorithmically derived, *i.e.*, can be computed for any chemical species directly from structure. The graph invariants used in this paper fall into this category. The results reported here show that such methods can be used as a first order estimation of properties.

The parameters used in this paper did not include any stereoelectronic property that might influence the normal boiling points of CFCs. It would be interesting to see whether similarity methods give better estimates of boiling points when stereoelectronic variables are included in the set of parameters. Such studies are in progress and will be reported subsequently.

Acknowledgments. — This paper is the contribution number 150 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported, in part, by grant F-49620-94-1-0401 from the United States Air Force, Exxon Biomedical Sciences, Inc. and the Structure-Activity Relationship Consortium (SARCON) of the Natural Resources Research Institute at the University of Minnesota. The authors would also like to extend their thanks to Professor A. T. Balaban of the Polytechnic University of Bucharest, Romania, for his helpful discussions.

REFERENCES

1. E. Mitscherlich, *Abhandl. Akad. Wiss. Berlin*, (1819) 427.
2. I. Langmuir, *J. Am. Chem. Soc.* **41** (1919) 1543.
3. T. Moeller, *Inorganic Chemistry*, John Wiley & Sons, New York 1852.
4. A. Cornish-Bowden and C. W. Wharton, *Enzyme Kinetics*, D. Rickwood (Ed.), IRL Press, Oxford, UK, 1988.
5. P. Calabresi and B. A. Chabner, *Antineoplastic Agents*, in: *The Pharmacological Basis of Therapeutics*, A. G. Gilman, T. W. Rall, A. S. Nies, and P. Taylor (Eds.), Eighth Edition, Pergamon Press, New York, 1990.
6. D. Voet and J. G. Voet, *Biochemistry*, John Wiley & Sons, New York, 1990.
7. J. Couraud, E. Escher, D. Regoli, V. Imhoff, B. Rossignol, and P. Pradelles, *J. Bio. Chem.* **260** (1985) 9461.
8. R. E. Carhart, D. H. Smith, and R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **25** (1985) 64.
9. C. L. Wilkins and M. Randić, *Theor. Chim. Acta (Berl.)* **58** (1980) 45.
10. D. H. Rouvray, *The evolution of the concept of molecular similarity*, in: *Concepts and Applications of Molecular Similarity*, M. A. Johnson and G. M. Maggiora (Eds.), John Wiley & Sons, New York, 15-42 (1990).
11. M. S. Lajiness, *Molecular similarity-based methods for selecting compounds for screening*, in: *Computational Chemical Graph Theory*, D. H. Rouvray (Ed.), Nova, New York, 299-316 (1990).
12. *Concepts and Applications of Molecular Similarity*, M. A. Johnson and G. M. Maggiora (Eds.), John Wiley & Sons, New York, 1990.
13. S. C. Basak, V. R. Magnuson, G. J. Niemi, and R. R. Regal, *Discrete Appl. Math.* **19** (1988) 17.
14. M. Johnson, S. C. Basak, and G. Maggiora, *Math. and Comp. Modelling* **II** (1988) 630.
15. S. C. Basak, S. Bertelsen, and G. Grunwald, *J. Chem. Inf. Comput. Sci.* **34** (1994) 270.
16. S. C. Basak and G. D. Grunwald, *SAR and QSAR Environ. Res.* **2** (1994) 289.
17. S. C. Basak and G. D. Grunwald, *New. J. Chem.* **19** (1995) 231.
18. S. C. Basak and G. D. Grunwald, *J. Chem. Inf. Comput. Sci.* **35** (1995) 366.
19. S. C. Basak and G. D. Grunwald, *SAR and QSAR Environ. Res.* **3** (1995) 265.
20. S. C. Basak and G. D. Grunwald, *Chemosphere* **31** (1995) 2529.
21. S. C. Basak, S. Bertelsen and G. D. Grunwald, *Toxicol. Lett.* **79** (1995) 239.
22. S. C. Basak and G. D. Grunwald, *Math. Modelling and Sci. Comput.*, in press.
23. P. Willet and V. Winterman, *Quant. Struct.-Act. Relat.* **5** (1986) 18.
24. W. Fisanick, K. P. Cross, and A. Rusinko, III, *J. Chem. Inf. Comput. Sci.* **32** (1992) 664.
25. P. E. Bowen-Jenkins, D. L. Cooper, and W. G. Richards, *J. Phys. Chem.* **89** (1985) 2195.
26. J. C. Arcos, *Environ. Sci. Tech.* **21** (1987) 743.
27. C. M. Auer, J. V. Nabholz, and K. P. Baetcke, *Environ. Health Perspect.* **87** (1990) 183.
28. C. M. Auer, M. Zeeman, J. V. Nabholz, and R. G. Clements, *SAR and QSAR Environ. Res.* **2** (1994) 29.
29. N. Trinajstić, *Chemical Graph Theory*, Second Edition, CRC Press, Inc., Boca Raton, Florida, 1992.
30. A. T. Balaban, S. C. Basak, T. Colburn, and G. D. Grunwald, *J. Chem. Inf. Comput. Sci.* **34** (1994) 1118.
31. A. T. Balaban, N. Joshi, L. B. Kier, and L. H. Hall, *J. Chem. Inf. Comput. Sci.* **32** (1992) 233.

32. S. C. Basak, D. K. Harriss, and V. R. Magnuson, POLLY: Copyright of the University of Minnesota, 1988.
33. H. J. Wiener, *J. Am. Chem. Soc.* **69** (1947) 17.
34. M. Randić, *J. Am. Chem. Soc.* **97** (1975) 6609.
35. L. B. Kier and L. H. Hall, *Molecular Connectivity in Structure-Activity Analysis*, Research Studies Press, Letchworth, Hertfordshire, UK 1986.
36. D. Bonchev and N. Trinajstić, *J. Chem. Phys.* **67** (1977) 4517.
37. C. Raychaudhury, S. K. Ray, J. J. Ghosh, A. B. Roy, and S. C. Basak, *J. Comput. Chem.* **5** (1984) 581.
38. A. B. Roy, S. C. Basak, D. K. Harriss, and V. R. Magnuson, *Neighborhood complexities and symmetry of chemical graphs, and their biological applications*, in: *Math. Modelling in Sci. and Tech.*, X. J. R. Avula, R. E. Kalman, A. I. Liapis, and E. Y. Rodin (Eds.), Pergamon Press, New York, 745-750 (1984).
39. S. C. Basak, A. B. Roy, and J. J. Ghosh, *Proc. IIInd. Int. Conf. on Math. Modelling* **II** (1980) 851.
40. S. C. Basak and V. R. Magnuson, *Arzneim.-Forsch.* **33** (1983) 501.
41. R. Sarkar, A. B. Roy, and P. K. Sarkar, *Math Biosci.* **39** (1978) 299.
42. A. T. Balaban, *Chem. Phys. Lett.* **89** (1982) 399.
43. A. T. Balaban, *Pure & Appl. Chem.* **55** (1983) 199.
44. A. T. Balaban, *MATCH* **21** (1986) 115.
45. SAS Institute Inc., In *SAS/STAT User's Guide, Release 6.03 Edition*, SAS Institute Inc., Cary, NC, Chapter 34 (1988) 949-965.

SAŽETAK

Procjena normalnih vrelišta haloalkana na osnovi molekulske sličnosti

Subhash C. Basak, Brian D. Gute i Gregory D. Grunwald

Molekulska sličnost upotrijebljena je za procjenu normalnih vrelišta skupa od 276 haloalkana s 1 do 4 ugljikova atoma. Molekulska sličnost/različitost kvantificirana je Euklidovom udaljenosću molekula u osmerodimensijском prostoru glavnih komponenti izvedenih iz 59 topoloških indeksa. Koeficijent korelacije između eksperimentalnih i procijenjenih vrelišta iznosi između 0.854 i 0.943 za procjene vrelišta pomoću K najbližih susjeda, uz različite brojeve najbližih susjeda ($K = 1, \dots, 10, 15, 20, 25$).

4

Use of Graph-Theoretic and Geometrical Molecular Descriptors in Structure-Activity Relationships

SUBHASH C. BASAK, GREGORY D. GRUNWALD, and
GERALD J. NIEMI

Ostensibly there is color, ostensibly sweetness, ostensibly bitterness, but actually only atoms and the void.
GALEN
(*Nature and the Greeks*, Erwin Schrödinger, 1954)

4.1. INTRODUCTION

One of the current interests in pharmaceutical drug design,¹⁻¹⁰ chemistry,²¹⁻⁴⁰ and toxicology⁴¹⁻⁵³ is the prediction of physicochemical, biomedicinal, and toxicological

SUBHASH C. BASAK, GREGORY D. GRUNWALD, and GERALD J. NIEMI • Center for Water and the Environment, Natural Resources Research Institute, University of Minnesota at Duluth, Duluth, Minnesota 55811.

From Chemical Topology to Three-Dimensional Geometry, edited by Balaban. Plenum Press, New York, 1997

properties of molecules from nonempirical structural parameters which can be calculated directly from their structure. Both in drug design^{3,4,31,33,54} and in hazard assessment of chemicals,^{31,33,46-53,55} one has to evaluate therapeutic or toxic potential of a large number of compounds, many of which have not even been synthesized. Drug design usually begins with the discovery of a "lead" compound which has the particular therapeutic activity of interest. The lead is altered through molecular modifications and the analogues thus produced are tested until a compound of desirable activity and toxicity profile is found. The combination of possibilities in such a process is almost endless. For example, let us assume the compound in Figure 1 is a lead. The medicinal chemist can carry out numerous manipulations on the lead in terms of substitution. On a very limited scale, if one carries out 50 substitutions in each of the aromatic positions, 10 modifications for esterification, 10 substitutions for the aliphatic carbon and 10 10 substitutions for the nitrogen, the total number of possible analogues comes to $50^5 \times 10 \times 10 \times 10 = 312.5$ billion structures. This astronomical number is reached by considering only a small fraction of the possible substituents that the medicinal chemist has in his repertoire.⁵⁴

A similar situation exists for the hazard assessment of environmental pollutants. More than 15 million distinct chemical entities have been registered with the Chemical Abstract Service and the list is growing by nearly 775,000 per year. About 1000 of these chemicals enter into societal use every year.⁵⁵ Few of these chemicals have experimental properties needed for risk assessment. Table 1 gives a partial list of properties necessary for a reasonable risk assessment of a chemical.^{31,33} In the United States, the Toxic Substances Control Act Inventory has about 74,000 entries and the list is growing by nearly 3000 per year. Of the approximately 3000 chemicals

Table 1. Properties Necessary for Risk Assessment of Chemicals

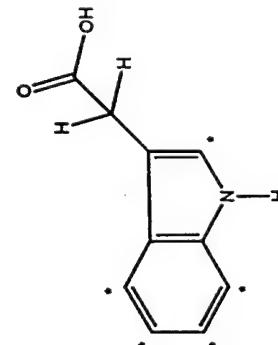
	Physicochemical	Biological
Molar volume	Receptor binding (K_D)	
Boiling point	Michaelis constant (K_m)	
Melting point	Inhibitor constant (K_I)	
Vapor pressure	Biodegradation ⁵⁶	
Aqueous solubility	Bioconcentration	
Dissociation constant (pKa)	Alkylation profile	
Partition coefficient	Metabolic profile	
Octanol-water ($\log P$)	Chronic toxicity	
Air-water	Carcinogenicity	
Sediment-water	Mutagenicity	
Reactivity (electrophile)	Acute toxicity	
	LD ₅₀	
	LC ₅₀	
	EC ₅₀	

submitted yearly to the U.S. Environmental Protection Agency for the premanufacture notification process, more than 50% have no experimental data, less than 15% have empirical mutagenicity data, and only about 6% have experimental ecotoxicological and environmental fate data.⁵⁵ Also, limited data are available for many of the over 700 chemicals found on the Superfund list of hazardous substances.

In the face of this massive unavailability of experimental data for the vast majority of chemicals, practitioners in drug discovery and hazard assessment have developed the use of nonempirical parameters to estimate molecular properties.^{1,3,4,20,31-33} By *nonempirical*, we mean those parameters that can be calculated directly from molecular structure without any other input of experimental data. Topological indexes (TIs), substructural parameters defined on chemical graphs, geometrical (3D or shape) parameters, and quantum-chemical parameters fall in this category.^{3,4,21-40,46-55,57-61}

A large number of quantitative structure-activity relationships (QSARs) pertaining to chemistry, pharmacology, and toxicology have used these nonempirical parameters. QSARs are mathematical models that relate molecular structure to their physicochemical, biomedical, and toxic properties. Two distinct processes are involved in the derivation of nonempirical parameters for a chemical: (1) defining the model object called "structure" which represents the salient features of the architecture of the chemical species and (2) calculating structural quantifiers from a selected set of critical features of the model object.^{31,62} Figure 2 depicts the process of experimental determination of properties vis-à-vis prediction of properties using descriptors.

Figure 2 represents an empirical property as a function $\alpha:C \rightarrow R$ which maps the set C of chemicals into the real line R. A nonempirical QSAR may be regarded as a composition of a description function, $\beta_1:C \rightarrow D$, mapping each chemical structure of C into a space of nonempirical structural descriptors (D) and a prediction function, $\beta_2:D \rightarrow R$, which maps the descriptors into the real line. When $[\alpha(C) - \beta_2\beta_1(C)]$ is within the range of experimental errors, we say that we have a good nonempirical



- 50 groups for each aromatic position (*)
- 10 groups for esterification
- 10 groups for aliphatic C
- 10 groups for ring N

$$\text{Total analogs} = 50^5 \times 10 \times 10 \times 10 = 312.5 \text{ billion}$$

Figure 1. Probable number of derivatives from a lead via molecular modification.

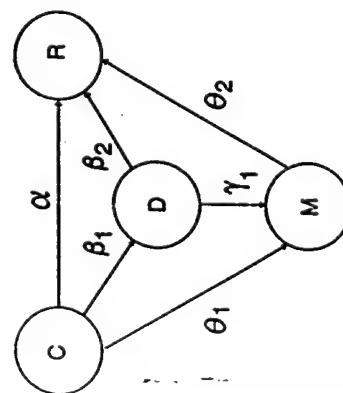


Figure 2. Composition functions for quantitative structure-activity relationship (QSAR) and property-activity relationship (PAR).

predictive model. On the other hand, a property-activity relationship (PAR) is the composition of $\theta_1:C \rightarrow M$, which maps the set C into the molecular property space M, and $\theta_2:M \rightarrow R$, mapping those molecular properties into the real line R. PAR seeks to predict one property (usually a complex property) of a molecule in terms of another (usually simpler or available) property. The latter group of properties may consist either of a number of experimentally determined quantities (e.g., melting point, boiling point, vapor pressure, partition coefficient) or substituent constants or solvatochromic parameters (e.g., steric, electronic, hydrophobic, charge transfer substituent constants, hydrogen bond donor acidity, hydrogen bond acceptor basicity).^{54,60} PAR using a calculated property, e.g., calculated partition coefficient ($\log P$, octanol-water), may be looked on as a mapping $\theta_2\gamma_1\beta_1:C \rightarrow R$, which is a composition of $\beta_1:C \rightarrow D$, $\gamma_1:D \rightarrow M$ mapping the descriptor space into the molecular property space (e.g., calculation of $\log P$ from fragments using additivity rule), and $\theta_2:M \rightarrow R$.

Graph invariants have been used in a large number of QSARs.¹⁻⁵³ A graph invariant is a graph-theoretic property that is preserved by isomorphism.^{63,64} A graph invariant may be a polynomial, a sequence of numbers, or a single numerical index. Numerical indexes derived from the topological characteristics of molecular graphs are called topological indexes. Molecular structures can be symbolized by graphs where the atomic cores are represented by vertices and covalent chemical bonds are depicted by edges of the graph. Such a graph depicts the connectivity of atoms in a chemical species irrespective of the metric parameters (e.g., equilibrium distance between nuclei, valence angles) associated with the molecular structure. It is in this sense that molecular graphs can be seen as topological, rather than geometrical, representations of molecular structure.⁶⁵ TIs are numerical quantifiers of molecular topology and are sensitive to such structural features of molecules as size, shape, symmetry, branching, and cyclicity. Two nonisomorphic graphs may have the same set of graph invariants. In that sense, TIs do not uniquely characterize molecular topology. Yet, it has to be emphasized that TIs quantify many salient aspects of molecular structure. As a result, different graph invariants have been successfully used in characterizing the structural similarity/dissimilarity of molecules.^{1-4,29,47,49,50,66}

4.2. TOPOLOGICAL INDEXES AND QSAR
 TIs have been used in developing QSAR models for predicting various properties. We give below some examples of successful QSARs using TIs. Definitions of the TIs used in the following equations and throughout this chapter may be found in Table 2.

Table 2. Symbols for Topological Indexes, Geometrical Parameters, and Hydrogen Bonding Parameter and Their Definitions

Index symbol	Definition
I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
T_B^W	Mean information index for the magnitude of distances
$\sqrt{T_B^W}$	Mean information index for the equality of distances
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
P	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
IC	Information content of the distance matrix partitioned by frequency of occurrences of distance h
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph neighborhood of vertices
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
IC_r	Mean information content or complexity of a graph based on the r^{th} ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r^{th} ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r^{th} ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
TIC_r	Total information content for r^{th} order neighborhood of vertices in a hydrogen-filled graph
h_{χ} or h_{χ_p}	Path connectivity index of order $h = 0-6$
h_{χ_C}	Cluster connectivity index of order $h = 3-6$
$h_{\chi_{CQ}}$	Chain connectivity index of order $h = 3-6$
$h_{\chi_{CC}}$	Path- c -cluster connectivity index of order $h = 4-6$
$h_{\chi_{CP}}$	Bonding path connectivity index of order $h = 0-6$
$h_{\chi_{CQ}}$	Bonding cluster connectivity index of order $h = 3-6$
$h_{\chi_{CCQ}}$	Bonding chain connectivity index of order $h = 3-6$
$h_{\chi_{CCP}}$	Bonding path- c -cluster connectivity index of order $h = 4-6$

(continued)

Table 2. (Continued)

Index symbol	Definition
$h\chi^v$	Valence path connectivity index of order $h = 0$ –6
$h\chi^C$	Valence cluster connectivity index of order $h = 3$ –6
$h\chi^CN$	Valence chain connectivity index of order $h = 3$ –6
$h\chi^PC$	Valence path–cluster connectivity index of order $h = 4$ –6
χ_t	Total structure index
$\Omega\text{-MCI}$	Orthogonal molecular connectivity indexes
τ	Branchedness indexes
κ	Shape indexes
ϕ	Flexibility indexes
A_3	Half-sum of the cube of the adjacency matrix
P_3	Polarity number: number of third neighbors
N_2	Gordon–Scantlebury index: number of second neighbors
P_h	Number of paths of length $h = 0$ –10
J	Balaban's J index based on distance
J^B	Balaban's J index based on multigraph bond orders
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii
$U\chi_X\chi_Y$	Balaban's information-based indexes on distance sums
AZV	Local vertex invariant based on the adjacency matrix, atomic numbers, and vertex degrees
D	Mean distance topological index for any graph
D_1	Mean distance topological index for acyclic graphs
Z	Hosoya index
HB_1	Hydrogen bonding potential of molecule
ID	Molecular identification numbers
V_W	Volume of molecule
$3D_{W,H}$	3D Wiener number including hydrogens
$3D_W$	3D Wiener number without hydrogens

4.2.1. Physicochemical Properties

4.2.1.1. Boiling Point of Alkanes

Needham *et al.*²¹ used TIs to develop a regression equation to predict the normal boiling point (BP) for 74 alkanes:

$$(1) \quad \text{BP} = -9.6 + 38.1(\chi) - 49.0(1/\chi) + 5.7(\chi_{PC}) - 94.5(\chi_d) + 8.4(\chi_p) \quad (N = 74, r = 0.999, s = 1.86, F = 9030)$$

Subsequently, Basak and Grunwald⁷⁸ derived the following equation:

$$(2) \quad \text{BP} = -263 + 237(\chi) + 18.6(\text{CIC}_2) \quad (N = 74, r = 0.997, s = 3.83, F = 5287)$$

4.2.1.2. Boiling Point of Chlorofluorocarbons (CFCs)

Balaban *et al.*⁷⁰ were able to model the boiling points of a large set of CFCs using TIs with the following equation:

$$(3) \quad \text{BP} = -73.65 + 33.21(1\chi^v - 0\chi^v) - 64.06(0\chi^v) + 94.46(1\chi) - 20.65(N_{Br}) - 22.18(N_I) + 6.36(2\chi^v - 1\chi^v) \quad (N = 532, r = 0.98, s = 10.94, F = 2953)$$

Using a backpropagation neural network (NN), Balaban *et al.*²⁵ successfully predicted BP for 276 CFCs. As inputs to the NN, the following parameters were used: J index, Wiener index (W), number of carbon atoms (N_c), number of chlorine atoms (N_{Cl}), and number of fluorine atoms (N_F). This NN resulted in a correlation (r) = 0.992 of observed BP with predicted BP, with a standard error (s) of 8.5°C. The data set used for NN model development consisted of 276 CFCs with, at most, four carbon atoms.

4.2.1.3. Lipophilicity of Diverse Sets of Compounds

Basak *et al.*²⁷ derived the following equation to predict lipophilicity ($\log P$, octanol–water):

$$(4) \quad \log P = 1.76 - 0.50(HB_1) - 5.28(1C_0) - 1.48(CIC_1) + 3.75(0\chi^v) + 0.41(1P_6) \quad (N = 382, r = 0.95, s = 0.27, F = 1186)$$

where HB_1 is a theoretically calculated hydrogen bonding parameter. Basak *et al.*³¹ developed a refined model for chemicals with HB_1 equal to zero:

$$(5) \quad \log P = -3.13 - 1.64(1C_0) + 2.12(0\chi_C) - 2.91(0\chi_{CH}) + 4.21(0\chi^v) + 1.06(4\chi^v) - 1.02(4\chi_{PC}) \quad (N = 137, r = 0.98, s = 0.26, F = 446)$$

4.2.1.4. Chromatographic Retention Time of Alkanes, Alkylbenzenes

Bonchev and Trinajstic⁷⁹ derived the following correlation for alkylbenzenes:

$$(6) \quad \text{RI} = 683 + 2.97(T_D^E) + 2.71(P_0 - 6) \quad N = 28, r = 0.99, s = 0.58$$

For alkanes, Kier and Hall⁷⁷ found the following relationship:

$$(7) \quad \text{RI} = -0.242 + 0.719(1\chi) + 0.125(3\chi_p) \quad (N = 18, r = 0.998, s = 0.045, F = 1702)$$

4.2.2. Biomedicinal Properties

4.2.2.1. Anesthetic Dose (AD_{50}) of Barbiturates

Basak *et al.*¹³ predicted AD_{50} of barbiturates using various TIs:

$$(8) \quad AD_{50} = -49.1 + 200(SIC_1) - 190(SIC_1)^2 \\ (N = 13, r = 0.76, s = 0.20, F = 6.6)$$

$$(9) \quad AD_{50} = -200 + 153(IC_1) - 28.3(IC_1)^2 \\ (N = 13, r = 0.74, s = 0.21, F = 6.1)$$

$$(10) \quad AD_{50} = -41.2 + 11.5(\chi) - 0.740(\chi)^2 \\ (N = 13, r = 0.72, s = 0.21, F = 5.4)$$

4.2.2.2. Analgesic Potency ($A\text{-ED}_{50}$) of Barbiturates

Basak *et al.*¹³ correlated $A\text{-ED}_{50}$ of barbiturates using graph-theoretic parameters:

$$(11) \quad A\text{-ED}_{50} = 4700 - 26300(SIC_1) + 36700(SIC_1)^2 \\ (N = 7, r = 0.97, s = 6.5, F = 29)$$

$$(12) \quad A\text{-ED}_{50} = 5280 - 2800(CIC_1) + 372(CIC_1)^2 \\ (N = 7, r = 0.96, s = 7.4, F = 27)$$

$$(13) \quad A\text{-ED}_{50} = 2400 - 444(\chi) + 20.4(\chi)^2 \\ (N = 7, r = 0.94, s = 9.1, F = 17)$$

4.2.2.3. Enzymatic Acetyl Transfer Reaction

Several TIs have been found to correlate with the enzymatic acetyl transfer reaction,¹² as shown by the following equations:

$$(14) \quad A_\chi = 3.20 - 0.62(\chi) \\ (N = 9, r = 0.88, s = 0.24, F = 23)$$

$$(15) \quad A_\chi = 2.67 - 0.83(IC_1) \\ (N = 9, r = 0.91, s = 0.20, F = 35)$$

$$(16) \quad A_\chi = 3.13 - 4.07(SIC_1) \\ (N = 9, r = 0.92, s = 0.20, F = 36)$$

4.2.2.4. Hill Reaction Inhibitory Potency of Triazinones⁶

$$(17) \quad P_{50} = -13.36 + 71.15(SIC_1) - 63.64(SIC_1)^2 \\ (N = 11, r = 0.937, s = 0.316, F = 28.6)$$

4.2.2.5. Complement Inhibition by Benzamidines⁸⁰

$$(18) \quad 1/\log_{10} C = -1.125 + 0.487(ID) + 0.011(O) \\ (N = 105, r = 0.941, s = 0.020, F = 391)$$

4.2.2.6. Binding of Barbiturates to Cytochrome P_{450}

Basak⁴³ used several TIs to correlate the binding of barbiturates to cytochrome P_{450} :

$$(19) \quad K_s = 27.79 - 36.78(IC_0) + 12.17(IC_0)^2 \\ (N = 10, r = 0.99, s = 0.01, F = 156.1)$$

$$(20) \quad K_s = 5.94 - 41.26(SIC_0) + 71.84(SIC_0)^2 \\ (N = 10, r = 0.99, s = 0.01, F = 224.3)$$

$$(21) \quad K_s = 35.74 - 18.45(H^D) + 2.38(H^D)^2 \\ (N = 10, r = 0.98, s = 0.01, F = 94.6)$$

4.2.3. Toxicological Properties

4.2.3.1. Nonspecific Narcotic Activity of Alcohols

Basak and Magnuson⁸¹ correlated the nonspecific narcotic activity (LC_{50}) of alcohols using TIs:

$$(22) \quad \log LC_{50} = 1.979 - 1.896(CIC_1) \\ (N = 10, r = 0.989, s = 0.323, F = 355.3)$$

4.2.3.2. Nonspecific Toxicity of Esters to *Pimephales promelas*⁴¹

$$(23) \quad \log LC_{50} = -0.774 - 0.364(CIC_1) - 0.774(\chi) \\ (N = 15, r = 0.965, s = 0.194, F = 81.1)$$

$$\log LC_{50} = 1.012 - 0.774(CIC_1) - 0.615(I_{D^W})$$

$$(24) \quad (N = 15, r = 0.961, s = 0.204, F = 72.7)$$

4.2.3.6. Inhibition of *p*-Hydroxylation of Aniline by Alcohols⁸⁴

$$(29) \quad pIC_{50} = -13.85 + 25.17(IC_0) - 27.89(SIC_1) - 1.87(CIC_2)$$

$$(N = 20, r = 0.96, F = 62.7)$$

*Basak et al.*⁴² correlated information- or complexity-based parameters with mutagenic potency of nitrosamines:

$$(25) \quad \ln R = 61.0 - 86.8(IC_0) + 29.2(IC_0)^2 \\ (N = 15, r = 0.96, s = 1.17, p < 0.001)$$

$$(26) \quad \ln R = 12.0 - 15.3(IC_1) + 3.84(IC_1)^2 \\ (N = 15, r = 0.98, s = 0.86, p < 0.001)$$

Table 3 gives more exhaustive information about the list of properties of different chemical classes that have been successfully correlated using TIs.

Table 3. Summary of QSARs Using Topological Parameters

Property	Chemical class	Variables ^b	Method	Citation	Ref. No.
BP	Aliphatic alcohols	MCI, ID, J, κ , Elec, LOV/LOIS	LR	Smeekens and Jurs	85
BP	Alkanes	MCI	NLR	Filip et al.	73
BP	Alkanes	χ_X	LR	Needham et al.	21
BP	Alkanes	$W, J, N_{Cl}, N_{Br}, N_F, N_I$	NN	Randic	35
BP	Haloalkanes			Balaban et al.	25
BP	Haloalkanes	MCI, N_X	LR	Balaban et al.	70
BP	Haloalkanes	N, MCl, κ, ϕ, J	LR	Balaban et al.	70
BP	Nonanes-dodecane	Z, W, p_3, N_3, A_3	LR	Gao and Hooya	86
BP	Paraffins	Platt's No.	LR	Platt	88
BP	Paraffins	W	LR	Wiener	87
CD	α -Amino acids	MCI	LR	Pogliani	76
CNDO/2 charge	Alkanes	MCI	LR	Hall and Kier	89
Cavity SA	Alcohols	LOV/LOIS	NLR	Filip et al.	73
d	Nonanes-dodecane	Z, W, p_3, N_2, A_3	LR	Gao and Hooya	86
d	Alkanes	LOV/LOIS	NLR	Filip et al.	73
d_4^{20}	Infinite linear polymers	W	LR, NLR	Mekenyan et al.	69
d_4^{20}	Organophosphorus	MCI	LR	Pogliani	90
d_C	Nonanes-dodecane	Z, W, p_3, N_3, A_3	LR	Gao and Hooya	86
ΔG_f	Nonanes-dodecane	Z, W, p_3, N_2, A_3	LR	Gao and Hooya	86
ΔH_{vp}	Alkanes	LOV/LOIS	NLR	Filip et al.	73
ΔS	Nonanes-dodecane	Z, W, p_3, N_2, A_3	LR	Gao and Hooya	86
Diverse profile	Diverse	TI	CCA	Boekkens and Niemi	91
E_t	Diverse	LOV/Sub.-TI	LR	Balaban and Catana	92
E_s	Hydrocarbons	χ_V	LR	Gupta and Singh	93
ΔH_F	Nonanes-dodecane	Z, W, p_3, N_2, A_3	LR	Gao and Hooya	86
ΔH_F	Paraffins	Platt's No.	LR	Platt	88

(continued)

4.2.3.3. Mutagenicity of Nitrosamines

*Basak et al.*⁴² correlated information- or complexity-based parameters with mutagenic potency of nitrosamines:

$$(25) \quad \ln R = 61.0 - 86.8(IC_0) + 29.2(IC_0)^2 \\ (N = 15, r = 0.96, s = 1.17, p < 0.001)$$

$$(26) \quad \ln R = 12.0 - 15.3(IC_1) + 3.84(IC_1)^2 \\ (N = 15, r = 0.98, s = 0.86, p < 0.001)$$

4.2.3.4. Mutagenicity of Diverse Structures

*Basak et al.*⁴⁶ used six TIs and four substructure (subgraph) indicator variables to develop a linear model to classify a set of 520 diverse chemicals as mutagens or nonmutagens as defined by the Ames mutagenicity test.⁸² The data set used in their study consisted of 260 mutagens and 260 nonmutagens. The TIs included three information-based indexes: information content of the graph orbits (*J*orB), information content at sixth order (*J*C₆), and structural information content at zeroth order (*SIC*₀).

A fourth index included number of paths of length 10 (*P*₁₀). The remaining two indexes were connectivity type: third-order bond-corrected cluster connectivity ($^3\chi_C$) and third-order valence-corrected chain connectivity ($^3\chi_h$). The four substructure indicators were: (1) nitroso chemicals, (2) halogen-substituted mustard, sulfur mustard, or oxygen mustard, (3) organic sulfate or sulfonate, and (4) a biphenyl amine, benzidine, or 4,4'-methylene dianiline derivative.

Using these parameters, a 74.8% overall correct classification rate was achieved. Jackknifed classification tests showed a 74.6% overall correct classification rate.

4.2.3.5. Toxicity of Monoketones

*Basak et al.*⁸³ derived the following correlations between TIs and the toxicity (*LD*₅₀) of monoketones:

$$(27) \quad LD_{50}(\text{control}) = 620.0 - 448.0(CIC_1) + 83.5(CIC_1)^2 \\ (N = 13, r = 0.95, s = 9.62, F = 48.9)$$

$$(28) \quad LD_{50}(\text{CCl}_4) = 407.0 - 235.0(CIC_0) + 35.1(CIC_0)^2 \\ (N = 13, r = 0.97, s = 4.76, F = 74.0)$$

Table 3. (Continued)

Property	Chemical class	Variables ^b	Method	Citation	Ref. No.
MON	Alkanes	J, D, D_1	LR	Balaban Filip <i>et al.</i>	23
MON	Alkanes	LOVILLOIS	NLR	Pogliani	73
MON	Alkanes	MCI	LR	Pogliani	90
MON	Alkanes	MCI	LR	Pogliani	90
MP	Caffeine	MCI	LR	Pogliani	90
MP	homologues	W	LR,NLR	Mekkenyan <i>et al.</i>	69
MP	Infinite linear polymers	Ω -MCI	LR	Randic	94
MR	Alkybenzenes	1st-order MCI	LR	Kupchik	75
MR	Alkybenzenes	Ω -MCI	LR	Randic	94
MR	Heptanes	Z, W, p_3, N_2, A_3	LR	Gao and Hosoya	86
MR	Nonanes—dodecanes	MCI	LR	Pogliani	90
MR	Organophosphorus	Platt's No.	LR	Platt	88
MR	Paraffins	Z, W, p_3, N_2, A_3	LR	Gao and Hosoya	86
MR	Nonanes—dodecanes	Z, W, p_3, N_2, A_3	LR	Platt	88
MR	Paraffins	MCI	LR	Pogliani	76
MR	α -Amino acids	Z, W, p_3, N_2, A_3	LR	Gao and Hosoya	86
MR	Nonanes—dodecanes	MCI	LR	Pogliani	90
MV	Organophosphorus	$J, X, Y, V, U,$ χ, AZV	LR	Balaban and Ferjou	74
MV	Alkanes	Z, W, p_3, N_2, A_3	LR	Gao and Hosoya	86
η_D	Nonanes—dodecanes	MCI	LR	Pogliani	76
η_D	α -Amino acids	ρ^*, P_0	LR	Kler and Hall	77
η_D	Alkanes	MCI	LR	Bonchev and Trinajstic	79
RI	Alkybenzenes	MCI	LR	Rohrbaugh and Jurs	95
RI	Diverse drugs	$MCI, P_{\alpha},$ $\kappa, Elec.$	LR	Pogliani	90
RI	Organophosphorus	MCI	LR	Pogliani Pal <i>et al.</i>	96
RI	Alkanes	τ	LR	Pogliani	76
RON	α -Amino acids	MCI	LR	Pogliani	90
S	Caffeine	MCI	LR	Pogliani	90
S	homologues	$J, X, Y, V, U,$ χ, AZV	LR	Balaban and Ferjou	74
T _C	Alkanes	Z, W, p_3, N_2, A_3	LR	Gao and Hosoya	86
T _C	Nonanes—dodecanes	W, J, MCI, ID	LR	Rouvray and Tatong	98
Ultrasound	Alkanes, alcohols	MCI	LR	Pogliani	76
V _P	α -Amino acids	W, J, MCI, N _{C1}	LR	Rouvray and Tatong	97
V _P	Polychlorinated biphenyls	$J, X, Y, V, U,$ χ, AZV	LR	Balaban and Ferjou	74
V _C	Alkanes	Z, W, p_3, N_2, A_3	LR	Gao and Hosoya	86
V _C	Nonanes—				

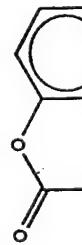
Table 3. (Continued)

Property	Chemical class	Variables ^b	Method	Citation	Ref. No.
$\alpha_{D,22}$	Infinite linear polymers	W	LR,NLR	Mektenyan <i>et al.</i>	69
$\log P$	Diverse	TI, HB ₁	LR	Basak <i>et al.</i>	27
$\log P$	Diverse	TI	LR	Niemi <i>et al.</i>	45
$\log P$	Diverse	MCI	LR,NLR, PCR	Niemi <i>et al.</i>	99
$\log P$	Diverse, HB ₁ = 0	TI	LR	Basak <i>et al.</i>	31
PI	α -Amino acids	MCI	LR	Pogliani	76
Bioactive	Bioactive agents	Inf.	LR	Ray <i>et al.</i>	9
Pharmacological	Benzamidines	Inf.	LR	Basak <i>et al.</i>	5
$1/\log C$	Barbiturates	TI	LR	Basak <i>et al.</i>	80
A-ED ₅₀	Barbiturates	MCI, Inf.	LR	Basak <i>et al.</i>	13
AD ₅₀	Barbiturates	MCI, Inf.	LR	Basak <i>et al.</i>	13
AD ₅₀	Barbiturates	MCI, Inf., W	NLR	Basak <i>et al.</i>	16
A _X	Anilines	MCI, Inf.	LR	Basak <i>et al.</i>	12
Antihistaminic	2-(Piperidin-4-ylamino)-1H-benzimidazoles	Sub.-W	LR	Lukovits	100
BOD	Diverse	MCI	Clustering, DA	Niemi <i>et al.</i>	99
BOD	Diverse	MCI, log ^P	Clustering, DA	Niemi <i>et al.</i>	101
BOD	Diverse	MCI, k, Sub.	DA	Gombar and Enslein	102
Carcinogenicity	Diverse	σ , k, MCI, Sub.	LR,DA	Blake <i>et al.</i>	103
Cytostatic activity	1 H-Isolindoleindoles	Sub.-W	LR	Lukovits	100
Estrogen binding	2-Phenylindoles	Sub.-W	LR	Lukovits	100
K _S	Barbiturates	TI	LR	Basak	43
LC ₅₀	Alcohols	CIC	LR	Basak and Magnusson	81
LC ₅₀	Esters	$W, 1_x, 1_x^y, \text{Inf.}$	LR	Basak <i>et al.</i>	41
LD ₅₀	Monoketones	TC ₀ , TIC ₁ , CIC ₀ , CIC ₁	NLR	Basak <i>et al.</i>	83
LD ₅₀	Diverse	σ , k, MCI, Sub.	LR,DA	Blake <i>et al.</i>	103
Mutagenicity	Sulfamates	W _r -paths	SIMCA	Okuyama <i>et al.</i>	104
Taste	Therapeutics	W _r -paths	Clustering	Randic	40
lnR	Nitrosamines	IC ₀ , IC ₁	NLR	Basak <i>et al.</i>	42
piC ₅₀	<i>N</i> -alkyl/norketobenzenones/ triazinones	Inf.	LR	Ray <i>et al.</i>	6
piC ₅₀	Alcohols	Inf.	LR	Magnusson <i>et al.</i>	84

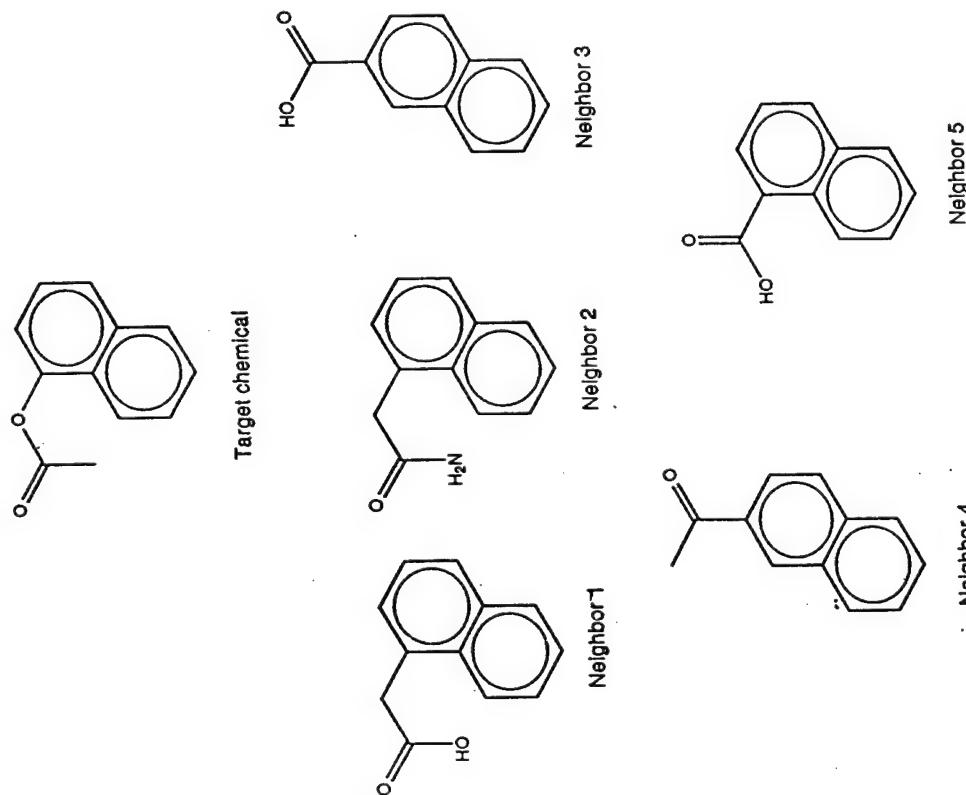
11

4.3. TOPOLOGICAL APPROACHES TO MOLECULAR SIMILARITY

One important application of TIs and substructural parameters has been in the quantification of molecular similarity. In practical drug design and risk assessment, good-quality QSARs of specific classes of chemicals, if available, are the best option. However, class-specific QSARs are often not available. In such cases, one selects analogues of the chemical of interest (lead or toxicant), and uses the property of



Target chemical



selected analogues for the estimation of the biomedicinal/toxic potential of the chemical.

4.3.1. Quantification of Similarity Using Path Numbers

Path numbers P_h ($h = 1, 2, \dots$) and weighted paths have been used by Randic and co-workers in determining partial orderings relating dopamine agonist properties for 2-aminotetralins, 105 physicochemical properties of decanes,¹⁰⁶ therapeutic potential of diverse compounds,⁴⁰ and antitumor activity of phenyldialkyltriazines.¹⁰⁷ Randic⁶⁶ has also reviewed the use of path numbers and weighted paths as they are applied in molecular similarity approaches to property optimization. The results show that the ordering of molecules by path numbers reflects the pattern of activity reasonably well.

4.3.2. Quantification of Similarity Using Topological Indexes

Basak *et al.*² used TIs to compute intermolecular similarity of chemicals. Ninety TIs were calculated for a set of 3692 chemicals with diverse structures. Principal component analysis (PCA) was used to reduce the 90-dimensional space to a 10-dimensional subspace which explained 93% of the variance. In the 10-dimensional PC space, the intermolecular similarity of chemicals were quantified in terms of their Euclidean distance (ED). Ten chemicals were then chosen at random from the set of 3692 structures and their analogues were selected using the Euclidean distance as the criterion for nearest-neighbor selection. Figure 3 gives one example of a probe chemical and its five chosen neighbors using this method. The results show that the probe and its selected analogues have a reasonable degree of structural similarity.

4.3.3. Quantification of Intermolecular Similarity Using Substructural Parameters

4.3.3.1. Atom Pairs (APs)

Carhart *et al.*⁴ developed the AP method of measuring molecular similarity. An AP is defined as a substructure consisting of two nonhydrogen atoms i and j and their interatomic separation:

$$\langle \text{atom descriptor}_i \rangle \langle \text{separation} \rangle \langle \text{atom descriptor}_j \rangle$$

where $\langle \text{atom descriptor}_i \rangle$ encodes information about the element type, number of nonhydrogen neighbors, and number of π electrons. Interatomic separation of two atoms is the number of atoms traversed in the shortest bond-by-bond path containing both atoms.

For two molecules, M_i and M_j , AP-based similarity is defined as:

$$(30) \quad S_{ij} = 2C/(T_i + T_j)$$

Figure 3. Target chemical and five selected analogues using ED method from the set of 3692 chemicals.

where C is the number of APs common to molecule i and j , T_i and T_j are the total number of APs in chemicals i and j , respectively. The numerator is multiplied by 2 to reflect the presence of shared APs in both molecules.

The Lederle group has used the AP similarity method to compare chemicals in their data base. Basak *et al.*^{28,29,46,47,49,50,53,108} have used the AP method in selecting analogues of chemicals in different and diverse data bases. The relative effectiveness of the AP and ED methods in selecting analogues of chemicals in the STARLIST¹⁰⁹ data base containing more than 4000 chemicals are shown in Figure 4.¹⁰⁸

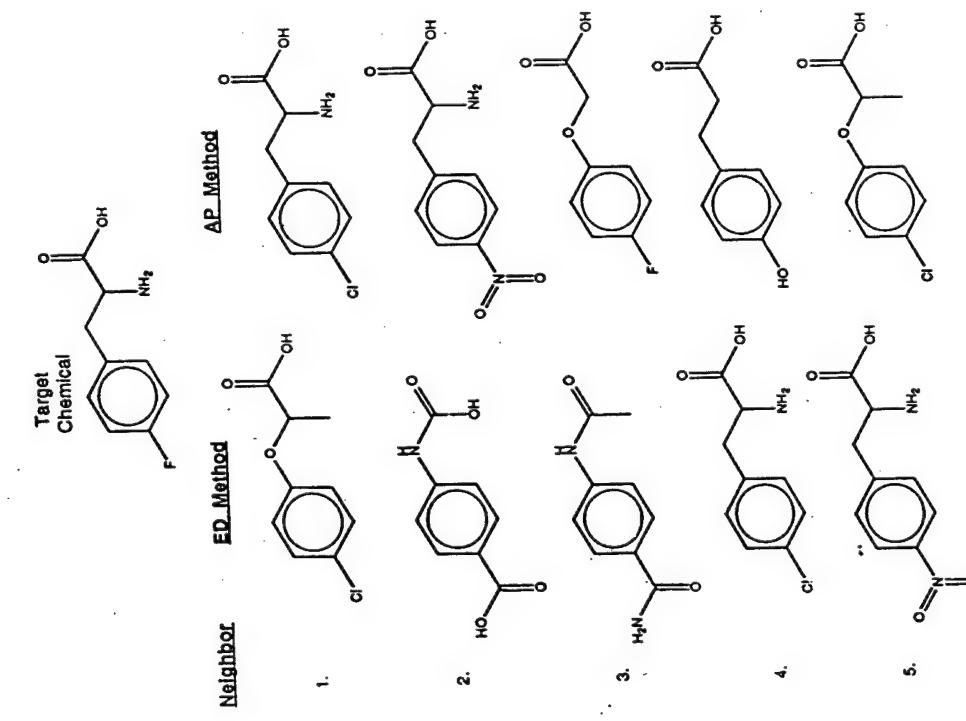


Figure 4. Target chemical and five selected analogues using ED and AP methods from the STARLIST data base of chemicals.

where C is the number of APs common to molecule i and j , T_i and T_j are the total number of APs in chemicals i and j , respectively. The numerator is multiplied by 2 to reflect the presence of shared APs in both molecules.

The Lederle group has used the AP similarity method to compare chemicals in their data base. Basak *et al.*^{28,29,46,47,49,50,53,108} have used the AP method in selecting analogues of chemicals in different and diverse data bases. The relative effectiveness of the AP and ED methods in selecting analogues of chemicals in the STARLIST¹⁰⁹ data base containing more than 4000 chemicals are shown in Figure 4.¹⁰⁸

4.3.3.2. Similarity Methods Based on Substructures

Willett and co-workers^{110–115} have developed several novel and useful techniques in molecular similarity based on substructural fragments. These approaches are based on the frequency of occurrence of generated fragment descriptors within the molecular graph. Success of these methods has been shown in 2D and 3D matchings of chemical structure, classification of chemical data bases, as well as property estimation.

4.3.4. *K* Nearest-Neighbor (KNN) Method of Estimating Properties

Basak and co-workers also used K ($K = 1–10, 15, 20, 25$) nearest neighbors of compounds in predicting properties like lipophilicity,²⁹ boiling point,^{28,47,49,116} and mutagenicity^{28,47,49,50} of diverse data bases. For a structurally diverse set of 76 compounds, lipophilicity ($\log P$, octanol–water) could be reasonably estimated using AP ($r = 0.85$) and ED ($r = 0.85$) methods for $K = 5$.²⁹

Four topologically based methods were used by Basak and Grunwald⁴⁷ in estimating the boiling point of a set of 139 hydrocarbons and a group of 15 nitrosoamines using the nearest neighbor ($K = 1$).

Basak and Grunwald⁴⁷ carried out a comparative study of five molecular similarity techniques, four topologically and one physicochemically based, in estimating the mutagenicity of a set of 73 aromatic and heteroaromatic amines. Of the five methods, two measures of molecular similarity were calculated using topological descriptors, two were derived using physical properties, and the fifth was based on a combination of both topological and physicochemical parameters. The best estimated values were obtained with $K = 4–5$.

Basak and Grunwald⁴⁹ also used topologically based similarity for KNN estimation of the mutagenicity of a set of 95 aromatic amines and the boiling point of a group of over 2900 chemicals with good results.

4.4. GEOMETRICAL/SHAPE PARAMETERS IN SAR

Geometrical parameters, such as molecular shape parameters,¹¹⁷ sterimol descriptors,⁵⁹ volume,⁶¹ bulk parameters,^{60,118} and 3D Wiener index,¹¹⁹ have been developed and used in SARs. Such parameters are derived from the relative distances of atoms in the 3D Euclidean space. We give below some examples of QSARs using 3D descriptors.

4.4.1. van der Waals Volume (V_w)

4.4.1.1. Physicochemical Properties

Bhatnagar *et al.*¹²⁰ studied the relationship of boiling point with V_w for several classes of chemicals, including saturated alcohols, primary amines, and alkyl halides:

$$(31) \quad \text{BP}_{\text{alcohols}} = 5.019 + 127.969(V_w) \\ (N = 48, r = 0.964, s = 8.25, F = 605)$$

$$(32) \quad \text{BP}_{\text{amines}} = -60.175 + 166.419(V_w) \\ (N = 21, r = 0.995, s = 5.13, F = 2061)$$

$$(33) \quad \text{BP}_{\text{alkyl halides}} = -108.431 + 226.874(V_w) \\ (N = 24, r = 0.896, s = 16.35, F = 90)$$

Correlation of water solubility (molarity) with V_w has also been determined for the saturated alcohols¹²⁰:

$$(34) \quad \log S = 6.908 - 8.596(V_w) \\ (N = 48, r = 0.974, s = 0.464, F = 860)$$

4.4.1.2. Biomedicinal Properties

Moriguchi and Kanada⁶¹ developed a regression equation modeling the effective concentration (C) of penicillins against *Staphylococcus aureus* in mice:

$$(35) \quad \log (1/C) = 5.911 - 1.692(V_w) \\ (N = 18, r = 0.927, s = 0.18)$$

4.4.1.3. Toxicological Properties

For tadpole narcosis of a diverse set of chemicals, the following equation has been developed⁶¹:

$$(36) \quad \log (1/C) = -2.022 + 2.940(V_w) \\ (N = 53, r = 0.969, s = 0.29)$$

Correlation of nonspecific toxicity on the Madison 517 fungus, expressed as $\log(1/C)$ (C is the minimum toxic dose), with V_w was found to be⁶¹:

$$(37) \quad \log (1/C) = -1.236 + 2.645(V_w) \\ (N = 45, r = 0.982, s = 0.19)$$

intersections of a 3D lattice. The partial least-squares method is used to describe statistical relationships between these fields and biological activity.

4.5. COMPARATIVE STUDY OF TOPOLOGICAL VERSUS GEOMETRICAL DESCRIPTORS IN QSARs

It is clear from the above that both topological and 3D descriptors have been extensively used in QSARs of large sets of molecules. However, no systematic work has been carried out on the relative effectiveness of TIs versus 3D parameters in the prediction of properties using QSAR models. We summarize below the results of our recent studies on the utility of graph-theoretic indexes and geometrical parameters such as 3D Wiener index and volume in estimating: (1) normal boiling point of a set of 140 hydrocarbons, (2) lipophilicity ($\log P$, octanol–water) of a diverse set of 254 molecules, and (3) mutagenic potency (in R , R being the number of revertants per manomole in the Ames test) of a set of 95 aromatic and heteroaromatic amines.

4.5.1. Property Data Bases

4.5.1.1. Boiling Point

All normal BP data for the hydrocarbons were found in the literature. The hydrocarbons analyzed include 74 alkanes,²¹ 29 alkyl benzenes,¹²² and 37 polycyclic aromatic hydrocarbons.¹²³ Table 4 presents a list of the hydrocarbon compounds with their normal BP (°C).

Table 4. Normal Boiling Point (°C) for 140 Hydrocarbons and Predicted Boiling Point Using Equations (44) and (45)

No.	Chemical name	Predicted BP		
		Obsd. BP	Eq. (44)	Eq. (45)
1	ethane	-88.6	-108.1	-94.7
2	<i>n</i> -propane	-42.1	-61.3	-47.7
3	<i>n</i> -butane	-0.5	-16.1	-2.3
4	2-methylpropane	-11.7	-17.9	-9.3
5	<i>n</i> -pentane	36.1	26.3	36.8
6	2-methylbutane	27.8	21.6	27.6
7	2,2-dimethylpropane	9.5	15.8	22.0
8	<i>n</i> -hexane	68.7	64.6	70.5
9	2-methylpentane	60.3	51.8	59.9
10	3-methylpentane	63.3	57.6	64.1
11	2,2-dimethylbutane	49.7	51.9	53.9
12	2,3-dimethylbutane	58.0	61.6	65.0
13	<i>n</i> -heptane	98.4	99.0	99.8
14	2-methylhexane	90.0	83.6	88.9

(continued)

Table 4. (Continued)

Table 4. (Continued)

No.	Chemical name	Obsd. BP		Predicted BP		Obsd. BP	Chemical name	Predicted BP	
		Eq. (44)	Eq. (45)	Eq. (44)	Eq. (45)			Eq. (44)	Eq. (45)
15	3-methylhexane	91.8	86.5	91.5	59	2,2,3-trimethylhexane	133.6	131.7	130.0
16	3-ethylpentane	93.5	91.6	98.3	60	2,2,4-trimethylhexane	126.5	130.2	125.4
17	2,2-dimethylpentane	79.2	75.6	80.3	61	2,2,5-trimethylhexane	124.1	129.0	124.1
18	2,3-dimethylpentane	89.8	90.9	91.1	62	2,3,3-trimethylhexane	137.7	134.8	132.9
19	2,4-dimethylpentane	80.5	83.6	89.3	63	2,3,4-trimethylhexane	139.0	144.9	140.7
20	3,3-dimethylpentane	86.1	81.8	83.6	64	2,3,5-trimethylhexane	131.3	138.5	137.4
21	2,2,3-trimethylbutane	80.9	88.4	88.4	65	2,4,4-trimethylhexane	130.6	133.0	128.0
22	<i>n</i> -octane	125.7	129.9	124.7	66	3,3,4-trimethylhexane	140.5	137.5	133.2
23	2-methylheptane	117.7	113.4	113.9	67	3,3-diethylpentane	146.2	145.0	144.5
24	3-methylheptane	118.9	115.1	116.2	68	2,2-dimethyl-3-ethylpentane	133.8	132.9	130.8
25	4-methylheptane	117.7	114.3	115.8	69	2,3-dimethyl-3-ethylpentane	142.0	140.0	136.7
26	3-ethylhexane	118.5	119.2	123.0	70	2,4-dimethyl-3-ethylpentane	136.7	144.9	139.2
27	2,2-dimethylhexane	106.8	102.7	104.3	71	2,2,3,3-tetramethylpentane	140.3	141.3	132.8
28	2,3-dimethylhexane	115.6	113.7	114.5	72	2,2,3,4-tetramethylpentane	133.0	139.3	134.7
29	2,4-dimethylhexane	109.4	114.9	112.2	73	2,2,4,4-tetramethylpentane	122.3	130.9	128.2
30	2,5-dimethylhexane	109.1	108.8	110.4	74	2,3,3,4-tetramethylpentane	141.6	145.0	139.2
31	3,3-dimethylhexane	112.0	105.2	106.4	75	benzene	80.1	99.2	76.0
32	3,4-dimethylhexane	117.7	119.0	117.9	76	toluene	110.6	121.3	112.2
33	2-methyl-3-ethylpentane	115.6	115.0	116.7	77	ethylbenzene	136.2	152.6	137.5
34	3-methyl-3-ethylpentane	118.3	111.8	115.6	78	<i>o</i> -xylene	144.4	142.3	146.7
35	2,2,3-trimethylpentane	109.8	111.7	109.1	79	<i>m</i> -xylene	139.1	137.3	135.1
36	2,2,4-trimethylpentane	99.2	106.5	105.1	80	<i>p</i> -xylene	138.4	137.3	134.5
37	2,3,3-trimethylpentane	114.8	115.0	112.3	81	<i>n</i> -propylbenzene	159.2	182.0	163.6
38	2,3,4-trimethylpentane	113.5	120.6	120.6	82	1-methyl-2-ethylbenzene	165.2	170.1	169.2
39	2,2,3,3-tetramethylbutane	106.5	119.0	112.7	83	1-methyl-3-ethylbenzene	161.3	166.8	158.7
40	<i>n</i> -nonane	150.8	158.0	147.3	84	1-methyl-4-ethylbenzene	162.0	166.0	157.9
41	2-methyloctane	140.8	136.7	136.7	85	1,2,3-trimethylbenzene	176.1	162.7	175.6
42	3-methyloctane	144.2	142.2	138.6	86	1,2,4-trimethylbenzene	169.4	161.4	166.5
43	4-methyloctane	142.5	140.7	138.4	87	1,3,5-trimethylbenzene	164.7	162.1	167.5
44	3-ethylheptane	143.0	145.4	145.8	88	<i>n</i> -butylbenzene	183.3	209.0	190.4
45	4-ethylheptane	141.2	144.7	145.5	89	1,2-diethylbenzene	183.4	195.2	192.8
46	2,2-dimethylheptane	132.7	128.7	126.2	90	1,3-dimethylbenzene	181.1	193.0	189.2
47	2,3-dimethylheptane	140.5	138.9	136.8	91	1,4-diethylbenzene	183.8	194.4	188.1
48	2,4-dimethylheptane	133.5	136.7	133.3	92	1-methyl-2- <i>n</i> -propylbenzene	184.8	194.8	193.5
49	2,5-dimethylheptane	136.0	137.5	133.8	93	1-methyl-3- <i>n</i> -propylbenzene	181.8	191.0	183.5
50	2,6-dimethylheptane	135.2	135.3	132.3	94	1-methyl-4- <i>n</i> -propylbenzene	183.8	190.4	182.5
51	3,3-dimethylheptane	137.3	130.4	128.5	95	1,2-dimethyl-3-ethylbenzene	193.9	187.4	196.2
52	3,4-dimethylheptane	140.6	141.3	139.1	96	1,2-dimethyl-4-ethylbenzene	189.8	186.6	187.5
53	3,5-dimethylheptane	136.0	142.8	138.8	97	1,3-dimethyl-2-ethylbenzene	190.0	187.2	193.5
54	4,4-dimethylheptane	135.2	130.2	127.9	98	1,3-dimethyl-4-ethylbenzene	183.4	186.4	190.4
55	2-methyl-3-ethylhexane	138.0	139.6	138.9	99	1,3-dimethyl-5-ethylbenzene	183.8	187.0	188.6
56	2-methyl-4-ethylhexane	133.8	137.2	136.7	100	1,4-dimethyl-2-ethylhexane	186.9	187.2	190.7
57	3-methyl-3-ethylhexane	140.6	135.9	136.5	101	1,2,3,4-tetramethylbenzene	205.0	185.2	202.9
58	3-methyl-4-ethylhexane	140.4	145.9	146.1	102	1,2,3,5-tetramethylbenzene	198.2	185.1	198.8

(continued)

4.5.1.2. Lipophilicity ($\log P$, Octanol–Water)

The 254 chemicals used to model $\log P$ are presented in Table 5. These chemicals were a subset of 382 chemicals studied by us previously²⁷ and consist of only those compounds with measured $\log P$ available in STARLIST,¹⁰ a selected subset of data deemed to be of very high quality by experts in the field. The $\log P$ values are provided in Table 5.

Table 4. (Continued)

No.	Chemical name	Obsd. BP	Predicted BP		Eq. (46)	Eq. (47)
			Eq. (44)	Eq. (45)		
103.	1,2,4,5-tetramethylbenzene	196.8	185.7	198.5		
104.	naphthalene	218.0	228.3	209.9		
105.	acenaphthalene	270.0	234.4	267.6		
106.	acenaphthene	279.0	—	272.0		
107.	fluorene	294.0	299.2	295.7		
108.	phenanthrene	338.0	346.4	329.8		
109.	anthracene	340.0	344.8	330.6		
110.	4H-cyclopenta(<i>de</i>)phenanthrene	359.0	326.9	349.7		
111.	fluoranthene	383.0	416.0	378.4		
112.	pyrene	393.0	392.2	384.3		
113.	benzo(<i>a</i>)fluorene	403.0	386.0	406.4		
114.	benzo(<i>b</i>)fluorene	398.0	390.8	406.4		
115.	benzo(<i>c</i>)fluorene	406.0	386.3	404.5		
116.	benzo(<i>gh</i>)fluoranthene	422.0	443.4	427.0		
117.	cyclopenta(<i>cd</i>)pyrene	439.0	—	435.7		
118.	chrysene	431.0	445.8	439.2		
119.	benz(<i>a</i>)anthracene	425.0	440.1	439.7		
120.	triphenylene	429.0	454.6	433.6		
121.	naphthacene	440.0	445.2	444.2		
122.	benzo(<i>b</i>)fluoranthene	481.0	503.9	476.0		
123.	benzo(<i>j</i>)fluoranthene	480.0	492.2	474.8		
124.	benzo(<i>k</i>)fluoranthene	481.0	501.1	489.9		
125.	benzo(<i>a</i>)pyrene	496.0	485.9	487.9		
126.	benzo(<i>e</i>)pyrene	493.0	490.8	484.4		
127.	perylene	497.0	484.0	479.7		
128.	anthanthrene	547.0	527.7	539.2		
129.	benzo(<i>gh</i>)perylene	542.0	526.9	529.7		
130.	indeno[1,2,3- <i>c,d</i>]fluoranthene	531.0	547.3	541.2		
131.	indeno[1,2,3- <i>c,d</i>]pyrene	534.0	540.1	534.5		
132.	dibenz(<i>a,c</i>)anthracene	535.0	535.2	533.7		
133.	dibenz(<i>a,h</i>)anthracene	535.0	528.9	546.2		
134.	dibenz(<i>a,f</i>)anthracene	531.0	529.6	543.6		
135.	picene	519.0	531.1	545.1		
136.	coronene	590.0	575.6	591.6		
137.	dibenz(<i>a,e</i>)pyrene	592.0	574.9	581.6		
138.	dibenz(<i>a,h</i>)pyrene	596.0	569.8	591.8		
139.	dibenz(<i>a,f</i>)pyrene	594.0	569.2	591.1		
140.	dibenz(<i>a,h</i>)pyrene	595.0	573.9	590.5		

Table 5. Observed $\log P$ and Estimated $\log P$ from Equations (46) and (47)

No.	Chemical name	Obsd. log P	Estimated log P for 254 Diverse Chemicals		Eq. (46)	Eq. (47)
			Estimated log P	Eq. (47)		
1.	butane	—	2.89	2.89	1.61	1.69
2.	pentane	—	3.39	3.39	2.08	2.16
3.	cyclopentane	—	3.00	3.00	2.61	2.46
4.	cyclohexane	—	3.44	3.44	2.27	2.46
5.	1-butene	—	4.00	4.00	1.65	1.57
6.	1-hexene	—	4.39	4.39	2.64	2.63
7.	cyclohexene	—	2.86	2.86	2.44	2.48
8.	1-pentyne	—	1.98	1.98	1.66	1.63
9.	ethylchloride	—	1.43	1.43	1.05	1.20
10.	1-chloropropane	—	2.04	2.04	1.43	1.65
11.	1-chlorobutane	—	2.64	2.64	1.81	2.04
12.	1-chloroheptane	—	4.15	4.15	3.09	3.32
13.	carbon tetrachloride	—	2.83	2.83	2.25	2.49
14.	1,2-dichloroethane	—	1.48	1.48	1.60	2.03
15.	1,1,1-trichloroethane	—	2.49	2.49	1.88	2.19
16.	1,1,2,2-tetrachloroethane	—	2.39	2.39	2.95	3.36
17.	trichloroethylene	—	2.83	2.83	2.86	3.16
18.	trichloroethylen	—	3.40	3.40	3.79	4.04
19.	trichlorofluoromethane	—	2.53	2.53	2.66	2.46
20.	benzene	—	2.13	2.13	2.21	2.14
21.	toluene	—	2.73	2.73	2.59	2.43
22.	<i>o</i> -xylene	—	3.12	3.12	3.09	2.98
23.	<i>m</i> -xylene	—	3.20	3.20	3.08	3.00
24.	<i>p</i> -xylene	—	3.15	3.15	2.89	2.79
25.	1,3,5-trimethylbenzene	—	3.42	3.42	3.61	3.58
26.	1,2,4-trimethylbenzene	—	3.78	3.78	3.62	3.57
27.	1,2,3-trimethylbenzene	—	3.66	3.66	3.60	3.50
28.	1,2,3,4-tetramethylbenzene	—	4.11	4.11	3.96	3.84
29.	1,2,3,5-tetramethylbenzene	—	4.17	4.17	4.25	4.14
30.	1,2,4,5-tetramethylbenzene	—	4.00	4.00	3.98	3.85
31.	pentamethylbenzene	—	4.56	4.56	4.70	4.73
32.	hexamethylbenzene	—	5.11	5.11	5.15	5.13
33.	ethylbenzene	—	3.15	3.15	2.94	2.83

(continued)

Table 5. (Continued)

No.	Chemical name	Obsd. log <i>P</i>	Estimated log <i>P</i>		Estimated log <i>P</i> Eq. (46)	Eq. (47)	Estimated log <i>P</i> Eq. (46)	Eq. (47)
			Eq. (46)	Eq. (47)				
34	propylbenzene	3.72	3.37	3.30	2.64	2.68	2.46	2.46
35	isopropylbenzene	3.66	3.35	3.30	2.75	2.87	2.68	2.68
36	butylbenzene	4.26	3.82	3.82	1.96	2.70	2.58	2.58
37	<i>t</i> -butylbenzene	4.11	3.76	3.71	2.30	3.00	2.95	2.95
38	<i>p</i> -cymene	4.10	3.85	3.78	1.83	2.69	2.53	2.53
39	fluorobenzene	2.27	2.37	2.00	82	3.36	3.32	3.32
40	chlorobenzene	2.84	2.37	2.42	83	2.73	3.68	3.70
41	bromobenzene	2.99	2.37	2.65	84	3.97	3.68	3.70
42	iodobenzene	3.25	2.37	3.15	85	1.49	2.32	2.09
43	<i>o</i> -dichlorobenzene	3.38	2.97	3.12	86	<i>o</i> -tolylacetate	2.89	2.72
44	1,3-dichlorobenzene	3.60	2.97	3.15	87	<i>m</i> -tolylacetate	2.09	2.61
45	<i>p</i> -dichlorobenzene	3.52	2.89	3.03	88	<i>p</i> -tolylacetate	2.11	2.66
46	1,2,3-trichlorobenzene	4.05	3.71	3.92	89	2-chlorophenyl acetate	2.18	2.59
47	1,2,4-trichlorobenzene	4.02	3.56	3.74	90	3-chlorophenyl acetate	2.32	2.62
48	1,3,5-trichlorobenzene	4.15	3.66	3.95	91	2-bromophenyl acetate	2.20	2.69
49	1,2,3,4-tetrachlorobenzene	4.64	4.34	4.55	92	propionaldehyde	0.87	0.64
50	1,2,3,5-tetrachlorobenzene	4.92	4.32	4.57	93	butyraldehyde	0.88	1.25
51	1,2,4,5-tetrachlorobenzene	4.82	4.31	4.56	94	hexaldehyde	1.78	2.56
52	Pentachlorobenzene	5.17	5.15	5.40	95	benzaldehyde	1.48	2.20
53	hexachlorobenzene	5.31	6.05	6.27	96	acetone	2.20	2.76
54	<i>o</i> -dibromobenzene	3.64	2.97	3.52	97	2-butane	0.24	0.47
55	<i>p</i> -dibromobenzene	3.79	2.89	3.48	98	2-pentanone	1.20	0.17
56	<i>o</i> -chlorotoluene	3.42	2.87	2.88	99	2-hexanone	0.29	0.17
57	<i>m</i> -chlorotoluene	3.28	2.87	2.90	100	2-heptanone	0.91	0.94
58	<i>p</i> -chlorotoluene	3.33	2.79	2.83	101	cyclohexanone	1.38	2.13
59	ethyl ether	0.89	1.23	1.21	102	acetophenone	1.48	2.13
60	dipropyl ether	2.03	2.12	2.15	103	<i>m</i> -chloroacetophenone	1.48	2.10
61	anisole	2.11	2.08	1.91	104	<i>p</i> -chloroacetophenone	1.98	2.58
62	diethyl ether	0.46	1.82	1.50	105	<i>p</i> -bromoacetophenone	0.81	2.58
63	tetrahydrofuran	1.04	1.25	1.09	106	<i>p</i> -fluoroacetophenone	1.58	2.52
64	ethyl vinyl ether	2.03	2.12	2.15	107	<i>p</i> -methylacetophenone	2.32	2.76
65	<i>o</i> -methylanisole	2.31	2.91	2.96	108	propiophenone	2.43	2.76
66	<i>m</i> -methylanisole	2.46	2.70	2.58	109	1-phenyl-2-propanone	1.44	2.95
67	<i>p</i> -methylanisole	2.78	2.36	2.39	110	ethylamine	1.72	2.76
68	4-chloroanisole	2.11	2.08	1.91	111	<i>m</i> -chloroacetophenone	2.51	2.89
69	phenetole	2.74	2.70	2.58	112	<i>p</i> -methylacetophenone	2.10	2.86
70	phenyl propyl ether	2.66	2.69	2.59	113	propiophenone	2.32	2.77
71	formic acid, propyl ester	2.81	2.59	2.50	114	1-phenyl-2-propanone	2.19	2.76
72	acetic acid, methyl ester	0.83	1.16	0.92	115	hexylamine	1.44	2.77
73	acetic acid, ethyl ester	0.18	0.72	0.18	116	heptylamine	0.48	0.66
74	propionic acid, ethyl ester	0.73	1.04	0.70	117	diethylamine	0.97	1.22
75	acrylic acid, methyl ester	0.80	1.03	0.50	118	dibutylamine	1.67	1.93
76	methacrylic acid, methyl ester	1.38	1.35	0.91	119	trimethylamine	0.81	2.70
77	benzoic acid, methyl ester	2.12	2.29	2.08	120	triethylamine	0.16	0.13

(continued)

Table 5. (Continued)

No.	Chemical name	Obsd. log <i>P</i>	Estimated log <i>P</i>		Eq. (46)	Eq. (47)	Eq. (46)	Eq. (47)
			Eq. (46)	Eq. (47)				
122	aniline	0.90	0.71	0.89			-0.01	0.19
123	<i>o</i> -toluidine	1.32	1.31	1.51			-0.77	0.61
124	<i>m</i> -toluidine	1.40	1.31	1.54			0.34	1.74
	<i>p</i> -toluidine	1.39	1.23	1.43			0.64	0.75
125	<i>m</i> -chloroaniline	1.88	1.12	1.46			-1.35	0.07
126	<i>m</i> -chloroaniline	1.83	1.04	1.38			1.95	1.72
127	<i>p</i> -chloroaniline	2.10	1.12	1.69			1.95	1.72
128	<i>m</i> -bromoaniline	2.26	1.04	1.62			-0.77	-0.13
129	<i>p</i> -bromoaniline	1.30	1.12	1.05			-0.77	-0.13
130	<i>m</i> -fluoroaniline	1.15	1.04	0.97			0.88	0.80
131	<i>p</i> -fluoroaniline	1.34	1.37	1.94			0.76	0.64
132	benzidine	2.25	2.20	2.39			1.56	1.30
133	α -naphthylamine	2.28	2.16	2.31			1.42	1.10
134	β -naphthylamine	2.31	2.45	2.47			2.03	1.71
135	<i>N,N</i> -dimethylaniline	2.81	2.95	2.96			2.97	2.45
136	<i>N,N</i> -dimethyl <i>p</i> -toluidine	3.31	3.40	3.44			0.17	0.38
137	<i>N,N</i> -diethylaniline	1.98	2.90	2.94			0.05	-0.00
138	<i>N,N</i> -dimethylbenzylamine	0.65	1.45	1.46			0.61	0.78
139	pyridine	1.20	1.69	1.65			1.21	1.04
140	3-methylpyridine	1.33	1.67	1.81			1.23	1.49
141	3-chloropyridine	1.60	1.67	2.07			0.35	0.26
142	3-bromopyridine	1.54	1.67	2.08			0.89	1.02
143	4-bromopyridine	-0.34	0.45	0.24			benzyl alcohol	1.59
144	acetonitrile	0.16	0.93	0.84			1.10	1.59
145	propionitrile	0.53	1.27	1.21			1.58	2.20
146	butyronitrile	1.56	2.17	2.01			1.94	2.10
147	benzonitrile	1.56	2.54	2.40			1.96	1.87
148	phenylacetonitrile	1.72	2.96	2.93			1.36	2.01
149	benzylacetonitrile	0.25	1.19	0.99			1.88	2.46
150	acrylonitrile	-0.35	-0.37	-1.01			1.95	2.36
151	nitromethane	0.18	0.47	0.10			1.46	1.28
152	nitroethane	0.87	0.75	0.52			1.96	1.84
153	1-nitropropane	1.47	1.21	1.03			1.94	1.75
154	1-nitrobutane	2.01	1.57	1.45			1.93	1.70
155	1-nitropentane	1.85	1.64	1.38			2.50	2.39
156	nitrobenzene	2.45	2.13	1.99			2.63	1.70
157	<i>m</i> -nitrotoluene	2.37	2.02	1.87			2.59	1.62
158	<i>p</i> -nitrotoluene	2.24	2.12	2.02			2.03	1.70
159	2-chloro-1-nitrobenzene	2.41	2.12	2.07			1.77	1.62
160	3-chloro-1-nitrobenzene	2.39	2.01	1.98			2.05	1.77
161	4-chloro-1-nitrobenzene	2.64	2.12	2.30			0.33	0.23
162	3-bromo-1-nitrobenzene	2.55	2.01	2.20			0.79	0.55
163	4-bromo-1-nitrobenzene	1.49	1.71	1.43			1.39	1.07
164	<i>m</i> -dinitrobenzene	1.46	1.63	1.38			1.92	1.43
165	<i>p</i> -dinitrobenzene						1.92	1.43

Table 5. (Continued)

Estimated log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Chemical name

No.

Obsd. log *P*

Eq. (46)

Chemical name

Eq. (46)

Eq. (47)

Obsd. log *P*

Eq. (46)

Eq. (47)

Table 5. (Continued)

No.	Chemical name	Obsd. log P	Estimated log P		Predicted lnR
			Eq. (46)	Eq. (47)	
210	decanoic acid	4.09	2.69	2.86	
211	benzoic acid	1.87	1.47	1.32	
212	<i>m</i> -toluic acid	2.37	2.01	1.95	
213	<i>p</i> -toluic acid	2.27	1.88	1.78	
214	<i>m</i> -chlorobenzoic acid	2.68	1.89	1.93	
215	<i>p</i> -chlorobenzoic acid	2.65	1.76	1.84	
216	<i>m</i> -bromobenzoic acid	2.87	1.89	2.13	
217	<i>p</i> -bromobenzoic acid	2.86	1.76	2.01	
218	<i>m</i> -fluorobenzoic acid	2.15	1.89	1.55	
219	<i>p</i> -fluorobenzoic acid	2.07	1.76	1.48	
220	phenylacetic acid	1.41	1.83	1.72	
221	<i>m</i> -chlorophenylacetic acid	2.09	2.13	2.23	1
222	<i>p</i> -chlorophenylacetic acid	2.12	2.12	2.23	2
223	<i>m</i> -bromophenylacetic acid	2.37	2.13	2.42	3
224	<i>o</i> -fluorophenylacetic acid	1.50	2.20	1.93	4
225	<i>m</i> -fluorophenylacetic acid	1.65	2.13	1.91	5
226	<i>p</i> -fluorophenylacetic acid	1.55	2.12	1.89	6
227	β -phenylpropionic acid	1.84	2.21	2.27	7
228	4-phenylbutyric acid	2.42	2.51	2.61	8
229	1-naphthoic acid	3.10	2.79	2.76	9
230	naphthalene	3.30	3.59	3.36	10
231	1-methylnaphthalene	3.87	4.07	3.91	11
232	2-methylnaphthalene	3.86	4.03	3.83	12
233	1,3-dimethylnaphthalene	4.42	4.53	4.39	13
234	1,4-dimethylnaphthalene	4.37	4.56	4.40	14
235	1,5-dimethylnaphthalene	4.38	4.46	4.28	15
236	2,3-dimethylnaphthalene	4.40	4.50	4.37	16
237	2,6-dimethylnaphthalene	4.31	4.46	4.32	17
238	1-nitronaphthalene	3.19	2.95	2.82	18
239	anthracene	4.45	4.80	4.59	19
240	9-methylanthracene	5.07	5.15	5.03	20
241	phenanthracene	4.46	4.83	4.66	21
242	pyrene	4.88	5.49	5.24	22
243	fluorene	4.18	3.69	3.77	23
244	acenaphthene	3.92	3.68	3.69	24
245	quinoline	2.03	2.69	2.60	25
246	isoquinoline	2.08	2.69	2.60	26
247	2,2'-biquinoline	4.31	4.49	4.63	27
248	biphenyl	4.09	4.10	3.97	28
249	2-chlorobiphenyl	4.38	4.27	4.24	29
250	2,4'-dichlorobiphenyl	5.10	4.56	4.62	30
251	2,5-PCB	5.16	4.59	4.68	31
252	2,6-PCB	4.93	4.67	4.72	32
253	2,4,6-PCB	5.47	4.99	5.09	
254	biphenyl	4.79	4.55	4.52	

(continued)

4.5.1.3. Mutagenicity (lnR)

The set of compounds used to model mutagenic potency consisted of 95 aromatic and heteroaromatic amines available from the literature.¹²⁴ A list of these chemicals and their mutagenic potency is presented in Table 6. The mutagenic potency of the aromatic amines in *S. typhimurium* TA98 + S9 microsomal preparation is expressed by the natural logarithm of the number of revertants per-nanomole.

Table 6. Mutagenicity (lnR)^a of 95 Aromatic and Heteroaromatic Amines and Predicted Mutagenicity by Equations (48) and (49)

No.	Chemical name	Obsd. lnR	Predicted lnR		
			Eq. (48)	Eq. (49)	Eq. (48)
1	2-bromo-7-aminofluorene	2.62	2.14	2.66	
2	2-methoxy-5-methylaniline	-2.05	-2.57	-2.07	
3	5-aminoquinoline	-2.00	-1.60	-1.71	
4	4-ethoxyaniline	-2.30	-3.75	-3.40	
5	1-aminoanthracene	-0.60	-0.93	-0.86	
6	4-aminofluorene	1.13	0.73	1.02	
7	2-aminonaphthalene	2.62	1.26	1.22	
8	7-aminofluoranthene	2.88	1.60	2.27	
9	8-aminoquinoline	-1.14	-1.79	-2.02	
10	1,7-diaminofluorene	0.75	0.18	0.23	
11	2-aminonaphthalene	-0.67	0.21	-0.42	
12	4-aminopyrene	3.16	2.99	2.89	
13	3-amino-3'-nitrobenzyl	-0.55	-0.26	0.19	
14	2,4,5-trimethylaniline	-1.32	-1.20	-0.55	
15	3-aminofluorene	0.89	1.35	1.37	
16	3,3'-dichlorobenzidine	0.81	0.24	0.95	
17	2,4-dimethylaniline	-2.22	-2.88	-2.34	
18	2,7-diaminofluorene	0.48	0.85	1.02	
19	3-aminofluoranthene	3.31	2.88	3.06	
20	2-aminofluorene	1.93	1.75	1.74	
21	2-aminoo-4'-nitrobenzyl	-0.62	0.06	0.20	
22	4-aminobiphenyl	-0.14	0.10	-0.04	
23	3-methoxy-4-methylaniline	-1.96	-3.21	-2.55	
24	2-aminocarbazole	0.60	0.61	0.25	
25	2-amino-5-nitrophenol	-2.52	-2.65	-3.16	
26	2,2'-diaminobiphenyl	-1.52	-0.42	-0.46	
27	2-hydroxy-7-aminofluorene	0.41	1.29	1.44	
28	1-aminophenanthrene	2.38	1.06	1.19	
29	2,5-dimethylaniline	-2.40	-2.41	-2.34	
30	4-amino-2'-nitrobenzyl	-0.92	0.06	0.09	
31	2-amino-4-methylphenol	-2.10	-3.59	-2.88	
32	2-aminophenazine	0.55	0.83	0.66	

Table 6. (Continued)

No.	Chemical name	Obsd. $\ln R$	Predicted $\ln R$		No.	Chemical name	Obsd. $\ln R$	Predicted $\ln R$	
			Eq. (48)	Eq. (49)				Eq. (48)	Eq. (49)
33	4-aminophenyl sulfide	0.31	0.32	0.18	77	1-amino-4-nitronaphthalene	-1.77	-0.21	-0.36
	2,4-dinitroaniline	-2.00	-0.59	-1.54	78	4-amino-3'-nitro biphenyl	1.02	-0.36	0.08
34	2,4-diaminoisopropylbenzene	-3.00	-1.79	-1.35	79	4-amino-4'-nitro biphenyl	1.04	-0.27	0.28
35	2,4-difluoroaniline	-2.70	-1.95	-2.68	80	1-aminophenazine	-0.01	0.67	0.51
36	2,4'-methyleneedianiline	-1.60	-1.23	-1.19	81	4,4'-methylenebis(α -fluoroaniline)	0.23	0.46	0.66
37	3,3'-dimethylbenzidine	0.01	-0.65	-0.76	82	4-chloro-2-nitroaniline	-2.22	-2.31	-2.79
38	2-aminofluoranthene	3.23	2.51	2.79	83	3-aminoquinoline	-3.14	-1.50	-1.82
39	2-amino-3'-nitro biphenyl	-0.89	-0.33	0.04	84	3-amino carbazole	-0.48	0.84	0.51
40	1-aminofluoranthene	3.35	2.61	2.92	85	4-chloro-1,2-phenylenediamine	-0.49	-1.50	-1.68
41	4,4'-ethylenbis(aniline)	-2.15	-1.79	-1.59	86	3-aminoanthracene	3.77	1.64	1.30
42	4-chloroaniline	-2.32	-2.54	-2.52	87	3,4'-diaminobiphenyl	0.20	-0.74	-1.03
43	2-aminophenanthrene	2.46	2.02	1.59	88	1-aminanthracene	1.18	1.32	1.33
44	4-fluoroaniline	-3.32	-2.85	-3.01	89	1-amino carbazole	-1.04	0.14	-0.11
45	9-aminophenanthrene	2.98	1.33	1.26	90	9-aminanthracene	0.87	1.65	1.45
46	3,3'-diaminobiphenyl	-1.30	-1.28	-1.14	91	4-amino carbazole	-1.42	0.33	0.06
47	2-aminopyrene	3.50	3.43	3.01	92	6-amino chrysene	1.83	2.49	3.31
48	2,6-dichloro-1,4-phenylenediamine	-0.69	-1.50	-1.78	93	1-aminopyrene	1.43	2.91	2.88
49	2-amino-7-acetamidoindole	1.18	1.09	1.46	94	4,4'-methylenebis(α -isopropylaniline)	-1.77	0.71	-1.29
50	2,8-diaminophenazine	1.12	0.17	0.34	95	2,7-diaminophenazine	3.97	1.42	1.10

$\ln R = \log$ revertants per nanomole, *S. typhimurium* TA98 with metabolic activation.

4.5.2. Calculation of Parameters

4.5.2.1. Computation of Topological Indexes

The first TI reported in the chemical literature, the Wiener index W ,⁸⁷ may be calculated from the distance matrix $D(G)$ of a hydrogen-suppressed chemical graph G as the sum of the entries in the upper triangular distance submatrix. The distance matrix $D(G)$ of a nondirected graph G with n vertices is a symmetric $n \times n$ matrix (d_{ij}), where d_{ij} is equal to the topological distance between vertices v_i and v_j in G . Each diagonal element d_{ii} of $D(G)$ is zero. We give below the distance matrix $D(G_1)$ of the labeled hydrogen-suppressed graph G_1 of isobutane (Figure 5):

$$D(G_1) = \frac{1}{2} \begin{bmatrix} 0 & 1 & 2 & 2 \\ 1 & 0 & 1 & 1 \\ 3 & 2 & 1 & 0 \\ 4 & 2 & 1 & 0 \end{bmatrix}$$

W is calculated as:

$$W = \frac{1}{2} \sum_{i,j} d_{ij} \cdot g_h$$

No.	Chemical name	Obsd. $\ln R$	Eq. (48)	Eq. (49)	No.	Chemical name	Obsd. $\ln R$	Eq. (48)	Eq. (49)
51	6-aminoquinoline	-2.67	-1.31	-1.51	52	4-methoxy-2-methylaniline	-3.00	-3.07	-2.55
53	3-amino-2'-nitro biphenyl	-1.30	-0.22	-0.10	54	2,4-diaminobiphenyl	-0.92	0.08	-0.31
55	1,6-diaminophenazine	0.20	0.41	0.54	56	4-aminophenyl disulfide	-1.03	-0.12	0.43
57	2-bromo-4,6-dinitroaniline	-0.54	-1.05	-1.33	58	2,4-diamino- <i>p</i> -butylbenzene	-2.70	-2.93	-3.09
59	4-aminophenyl ether	-1.14	-0.50	-0.56	60	2-aminobiphenyl	-1.49	0.02	-0.40
61	1,9-diaminophenazine	0.04	0.13	0.26	62	1-aminofluorene	0.43	0.99	1.11
63	8-aminofluoranthene	3.80	2.77	3.01	64	2-chloroaniline	-3.00	-3.16	-2.95
65	3-amino- α,α -trifluorotoluene	-0.80	-0.78	-1.16	66	2-amin-1-nitronaphthalene	-1.17	-0.24	-0.47
67	3-amino-4'-nitro biphenyl	0.69	-0.29	0.28	68	4-bromoaniline	-2.70	-2.23	-2.04
69	2-amino-4-chlorophenol	-3.00	-2.56	-2.68	70	3,3'-dimethoxybenzidine	-0.50	-0.43	-1.89
71	4-cyclohexylaniline	-1.24	-1.97	-0.80	72	4-phenoxylaniline	0.38	-0.28	-0.80
73	4,4'-methylenebis(α -ethylaniline)	-0.99	-0.11	-1.62	74	2-amino-7-nitrofluorene	3.00	1.56	2.36
75	Benzidine	-0.39	-0.98	-0.94	76				

To account for the chemical nature of vertices, as well as their bonding pattern, Sarkar, Roy, and Sarkar¹²⁷ calculated information content of molecular graphs on the basis of an equivalence relation where two atoms of the same element are considered equivalent if they possess an identical first-order topological neighborhood. Since properties of atoms or reaction centers are often modulated by physicochemical characteristics of distant neighbors, i.e., neighbors of neighbors, it was deemed essential to extend this approach to account for higher-order neighbors of vertices. This can be accomplished by defining open spheres for all vertices of a molecular graph. If r is any nonnegative real number, and v is a vertex of the graph G , then the open sphere $S(v, r)$ is defined as the set consisting of all vertices v_i in G such that $d(v, v_i) < r$. Obviously, $S(v, 0) = \phi$, $S(v, r) = v$ for $0 < r < 1$, and if $1 < r < 2$, then $S(v, r)$ is the set consisting of v and all vertices v_i of G situated at unit distance from v .

One can construct such open spheres for higher integral values of r . For a particular value of r , the collection of all such open spheres $S(v, r)$, where v runs over the whole vertex set V , forms a neighborhood system of the vertices of G . A suitably defined equivalence relation can then partition V into disjoint subsets consisting of topological neighborhoods of vertices up to r^{th} order neighbors. Such an approach has already been developed and the information-theoretic indexes calculated are called indexes of neighborhood symmetry.¹²⁸

In this method, chemical species are symbolized by weighted linear graphs. Two vertices v_0 and v_0 of a molecular graph are said to be equivalent with respect to r^{th} order neighborhood if and only if corresponding to each path u_0, u_1, \dots, u_r of length r , there is a distinct path v_0, v_1, \dots, v_r of the same length such that the paths have similar edge weights, and both u_0 and v_0 are connected to the same number and type of atoms up to the r^{th} order bonded neighbors. The detailed equivalence relation is described in our earlier studies.¹²⁸

Once partitioning of the vertex set for a particular order of neighborhood is completed, IC_r is calculated by equation (39). Basak, Roy, and Ghosh¹²⁹ defined another information-theoretic measure, structural information content (SIC_r), which is calculated as:

$$SIC_r = IC_r / \log_2 n \quad (40)$$

where IC_r is calculated from equation (39) and n is the total number of vertices of the graph.

Another information-theoretic invariant, complementary information content (CIC_r), is defined as⁸¹:

$$CIC_r = \log_2 n - IC_r \quad (41)$$

CIC_r represents the difference between maximum possible complexity of a graph (where each vertex belongs to a separate equivalence class) and the realized topological information of a chemical species as defined by IC_r . Figure 6 provides an example of the first order ($r = 1$) calculations of IC , SIC , and CIC .

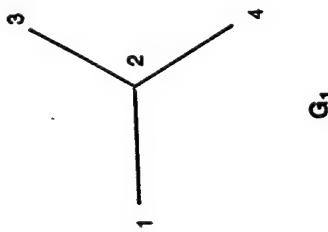


Figure 5. Labeled hydrogen-suppressed graph of isobutane.

where g_h is the number of unordered pairs of vertices whose distance is h . Randic's connectivity index,³⁵ higher-order connectivity indexes, and path, cluster, path-cluster, and chain types of simple, bond and valence connectivity parameters developed by Kier and Hall⁷⁷ were calculated by a computer program POLLY 2.3 developed by Basak, Harriss, and Magnuson¹²⁵ at the University of Minnesota. Also, developed by Basak, Harriss, and Magnuson¹²⁵ in the hydrogen-suppressed P_h parameters, the number of paths of length h ($h = 0-10$) in the hydrogen-suppressed graph, are calculated using standard algorithms.

Balaban²²⁻²⁴ defined a series of indexes based on distance sums within the distance matrix for a molecular graph which he designated as J indexes. Unlike W , these indexes are independent of molecular size and have low degeneracy. Information-theoretic TIs are calculated by the application of information theory on molecular graphs. An appropriate set A of n elements is derived from a molecular graph G depending on certain structural characteristics. On the basis of an equivalence relation defined on A , the set A is partitioned into disjoint subsets A_i of order n_i ($i = 1, 2, \dots, h$; $\sum_i n_i = n$). A probability distribution is then assigned to the set of equivalence classes:

$$A_1, A_2, \dots, A_h$$

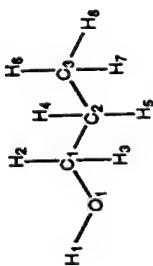
$$p_1, p_2, \dots, p_h$$

where $p_i = n_i/n$ is the probability that a randomly selected element of A will occur in the i^{th} subset. The mean information content of an element of A is defined by Shannon's relation¹²⁶:

$$IC = - \sum_{i=1}^h p_i \log_2 p_i \quad (39)$$

The logarithm base 2 is used to measure the information content in bits. The total information content of the set A is then n times IC.

theoretic parameters defined on the distance matrix, H^D and H^N , were calculated by the method of Raychaudhury *et al.*³⁶



4.5.2.2. Computation of Geometrical Parameters

Volume (V_w) was calculated using the SYBYL¹³⁰ package from Tripos Associates, Inc. The 3D Wiener numbers were calculated using SYBYL with an SPL (Sybyl Programming Language) program. Calculation of 3D Wiener numbers consists of the sum of entries in the upper triangular submatrix of the topographic Euclidean distance matrix for a molecule. The 3D coordinates for the atoms were determined using CONCORD 3.01.¹³¹ Two variants of the 3D Wiener number were calculated. For $3DW_H$, hydrogen atoms are included in the computations and for $3DW$, hydrogen atoms are excluded from the computations.

4.5.2.3. Computation of HB_1

The hydrogen bonding parameter HB_1 was calculated using a program developed by Basak.¹³² This program is based on the ideas of Ou *et al.*¹³³ The list of parameters used in this chapter is given in Table 2.

Subsets:	I	II	III	IV	V	VI
	(H ₁)	(H ₂ -H ₈)	(O ₁)	(C ₁)	(C ₂)	(C ₃)
Probability:						
I			III	IV	V	VI
1/12	7/12	1/12	1/12	1/12	1/12	1/12

4.5.3. Statistical Methods

4.5.3.1. Index Selection

Since the scale of the TIs vary by several orders of magnitude, each TI was transformed by the natural log of the index plus one.

The large number of TIs, and the fact that many of them are highly correlated, confounds the development of predictive models. Therefore, we attempted to reduce the number of TIs to a smaller set of relatively independent variables. Variable clustering¹³⁴ was used to divide the TIs into disjoint subsets (clusters) that are essentially unidimensional. These clusters form new variables which are the first principal component derived from the members of the cluster. From each cluster of indexes, a single index was selected. The index chosen was the one most correlated with the cluster variable. In some cases, a member of a cluster showed poor group membership relative to the other members of the cluster, i.e., the correlation of an index with the cluster variable was much lower than the other members. Any variable showing poor cluster membership was selected for further studies as well. A correlation of a TI with the cluster variable less than 0.7 was used as the definition of poor cluster membership.

$$I_D^W = W \log_2 W - \sum_h g_h \cdot h \log_2 h$$

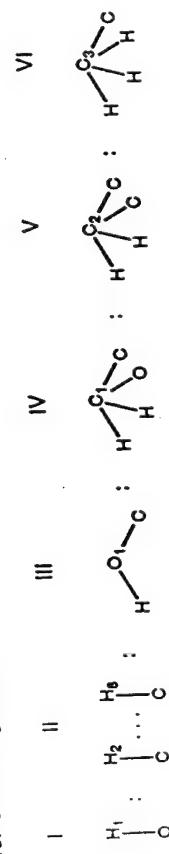
4.5.3.2. Regression Analysis

The variables used to model each of the properties in this study were TIs, HB_1 , and three geometry-related parameters, volume (V_w) and the two 3D Wiener numbers W , IC_1 , SIC_1 , CIC_1 , IC_r , SIC_r , and CIC_r were calculated by POLLY 2.3.¹²⁵ The information-

Figure 6. Derivation of first-order neighborhoods and calculation of complexity indexes (IC_1 , SIC_1 , and CIC_1) for n-propanol.
The information-theoretic index on graph distance, I_D^W is calculated from the distance matrix $D(G)$ of a molecular graph G as follows³⁴:

$$\begin{aligned} IC_1 &= 5 \cdot 1/12 \cdot \log_2 12 + 7/12 \cdot \log_2 12/7 = 1.950 \text{ bits} \\ SIC_1 &= IC_1 / \log_2 12 \\ CIC_1 &= \log_2 12 - IC_1 \end{aligned}$$

Figure 6. Derivation of first-order neighborhoods and calculation of complexity indexes (IC_1 , SIC_1 , and CIC_1) for n-propanol.



${}^3\text{DW}$ and ${}^3\text{D}W_{\text{H}}$). The TIs were restricted to those selected by the variable clustering procedure described previously.

All subsets regression was used for the development of the models. The criteria used for defining the “best” model were R^2 and Mallow’s C_p .¹³⁵ For each of the properties examined, initial models used only the TIs and HB_1 as potential variables. Subsequently, we added the three geometric variables to examine the improvement provided by the addition of geometric information.

4.5.4. Results

4.5.4.1. Boiling Point

HB_1 is zero for all hydrocarbons and, therefore, was deleted from analyses of BP. Twelve of the TIs were deleted for the analysis of the 140 hydrocarbons as well. These indexes included the third- and fourth-order chain connectivity indexes, which were zero for all chemicals, the fourth- and sixth-order bond and valence corrected cluster connectivity indexes, which were perfectly correlated with the simple cluster connectivity indexes ($r = 1.0$), and γ^X and γ^Y , which were perfectly correlated with γ^B for hydrocarbons.

Variable clustering of the remaining 89 TIs resulted in ten clusters. These clusters explained 89.7% of the total variation. In Table 7, we present the indexes selected from each cluster for subsequent use in modeling the BP of hydrocarbons. O , IC_0 , IC_1 , IC_2 , SIC_0 , and SIC_1 were selected because of their poor relationship with their clusters ($r < 0.7$).

With the 16 TIs, all subsets regression resulted in a seven-parameter model as follows:

$$\text{BP} = -322.86 + 5.46(P_4) - 45.76(\text{IC}_1) - 53.23(\text{IC}_2) + 799.94(\delta\chi_{\text{CH}}) +$$

Table 7. Topological Indexes Selected by Variable Clustering 89 Indexes for the Set of 140 Hydrocarbons with Measured Boiling Point

Cluster	Indices selected	Correlation
1	P_4	0.992
2	CIC_4 , IC_0	0.968, 0.338
3	IC_3 , O , IC_2 , SIC_0	0.959, 0.400, 0.581, 0.456
4	$\delta\chi_{\text{CH}}^Y$	0.964
5	$\delta\chi_{\text{CH}}^X$	0.998
6	$\delta\chi_{\text{CH}}^Y$, SIC_1	0.986, 0.537
7	$\delta\chi_{\text{CH}}^X$	0.990
8	$\delta\chi_{\text{CH}}^Y$, IC_1	0.916, 0.292
9	CIC_2	0.986
10	$\delta\chi_{\text{CH}}^X$	0.973

(${}^3\text{DW}$ and ${}^3\text{D}W_{\text{H}}$). The TIs were restricted to those selected by the variable clustering procedure described previously.

All subsets regression was used for the development of the models. The criteria used for defining the “best” model were R^2 and Mallow’s C_p .¹³⁵ For each of the properties examined, initial models used only the TIs and HB_1 as potential variables. Subsequently, we added the three geometric variables to examine the improvement provided by the addition of geometric information.

$$(43) \quad 288.26(\delta\chi^Y) - 32.76(\delta\chi^Y_{\text{CH}}) - 2518.52(\delta\chi^Y_{\text{CH}}) \\ (N = 140, r = 0.9956, s = 15.5, F = 2114)$$

Two chemicals, acenaphthene and cyclopenta(*cis*)pyrene (Nos. 106 and 117 of Table 4, respectively) had rather large residuals ($> 60^\circ \text{C}$). The Cook’s distance¹³⁶ for these two chemicals indicated they were influential cases. Given these circumstances, an outlier test¹³⁶ was performed and both chemicals had a significant result. After the removal of these chemicals, the following model was developed:

$$\text{BP} = -349.11 - 0.71(P_4) - 31.93(\text{IC}_1) - 44.70(\text{IC}_2) + 884.75(\delta\chi_{\text{CH}}) +$$

$$(44) \quad 291.24(\delta\chi^Y) - 33.10(\delta\chi^Y_{\text{CH}}) - 3327.61(\delta\chi^Y_{\text{CH}}) \\ (N = 138, r = 0.9976, s = 11.4, F = 3876)$$

With the inclusion of the geometric parameters, an eight-parameter model was developed, which included two of the geometric parameters:

$$\text{BP} = -626.4 + 1050.8(\text{SIC}_0) - 204.0(\text{SIC}_1) - 249.8(\delta\chi_{\text{CH}}) + 364.0(\delta\chi^Y) -$$

$$(45) \quad 32.3(\delta\chi^Y_{\text{CH}}) + 833.4(\delta\chi^Y_{\text{CH}}) + 20.4(\delta\text{DW})^{1/2} - 8.0(\delta\text{DW}_{\text{H}})^{1/2} \\ (N = 140, r = 0.9994, s = 6.1, F = 12246)$$

Table 4 presents the predicted normal BP for the hydrocarbons when using equations (44) and (45).

4.5.4.2. Lipophilicity ($\log P$)

Twelve of the TIs were dropped from the study of the $\log P$ data set. The third- and fourth-order chain connectivity indexes were zero for all chemicals and the fifth-order chain connectivity index was nonzero for only one chemical. The sixth-order cluster connectivity indexes were nonzero for only one compound as well. Therefore, 89 indexes were used for the variable clustering.

There were 12 clusters generated by variable clustering for the 89 TIs used for the $\log P$ data set. The total variation explained by these clusters was 87.8% of the total. Table 8 presents the indexes selected from each of the clusters. The indexes O , SIC_0 , J , IC_0 , IC_1 , and SIC_2 showed poor membership ($r < 0.7$) within the clusters and were retained as well.

Using all subsets regression with the selected TIs and HB_1 as independent variables resulted in a nine-parameter model:

$$\log P = -3.64 + 4.81(P_0) - 0.54(\text{IC}) - 9.30(\text{IC}_0) + 13.65(\text{SIC}_0) + \\ 3.88(\text{SIC}_4) - 7.68(\delta\chi_{\text{CH}}) + 0.63(\delta\chi_{\text{CH}}^Y) - 1.52(\delta\chi_{\text{CH}}^B) - 0.49(\text{HB}_1) \\ (N = 254, r = 0.912, s = 0.56, F = 134.1)$$

Table 8. Topological Indexes Selected by Variable Clustering 89 Indexes for the Set of 254 Chemicals with Measured $\log P$

Cluster	Indices selected	Correlation
1	P_0	0.981
2	SIC_4	0.944
3	χ_{PC}^b	0.929
4	IC_3, SIC_0, O	0.972, 0.469, 0.681
5	χ_{PC}^b	0.976
6	χ_C^b	0.980
7	χ_{CH}^b, J	0.855, 0.406
8	SIC_1, IC_0, IC_1, SIC_2	0.910, 0.503, 0.535, 0.689
9	J^b	0.996
10	χ_{PC}^b	0.988
11	IC	0.843
12	χ	0.963

Inclusion of the geometric parameters resulted in the following 11-parameter model:

$$\begin{aligned}
 \log P = & -12.06 - 0.68(\overline{IC}) - 8.13(IC_0) + 2.25(IC_3) + 12.62(SIC_0) - \\
 (47) \quad & 5.65(\chi_{CH}^b) + 0.66(\chi_{PC}^b) - 2.22(J) - 0.37(HB_1) + \\
 & 4.23 \log(V_w) + 0.60 \log(^3D W) - 0.75 \log(^3D W_H) \\
 (N = 254, r = 0.932, s = 0.50, F = 129.1)
 \end{aligned}$$

Predicted values of $\log P$ using equations (46) and (47) are presented in Table 5.

4.5.4.3. Mutagenicity

Twelve TIs were dropped from the analyses of the 95 aromatic and heteroaromatic amines. The indexes dropped included the third- and fourth-order chain connectivity indexes, which were zero for all chemicals, and the fourth- and sixth-order cluster connectivity indexes, which were nonzero for only three compounds. There were eight clusters generated by variable clustering the 89 TIs used for the aromatic amine data set. The total variation explained by these clusters was 88.6% of the total. In Table 9, we present the indexes selected from each of the clusters. Indexes $O, IC, IC_2, SIC_3, CIC_3$, and χ_C^b were retained because of their poor cluster membership ($r < 0.7$).

Using all subsets regression with the selected TIs and HB_1 as independent variables resulted in an eight-parameter model:

$$\ln R = 9.308 + 5.141(\overline{IC}) - 3.018(O) - 23.814(IC_3) - 15.050(SIC_1) +$$

Table 9. Topological Indexes Selected by Variable Clustering 89 Indexes for the Set of 95 Aromatic and Heteroaromatic Amines with Measured $\ln R^b$

Cluster	Cluster	Index selected	Correlation
1	1	χ	0.982
2	2	SIC_4, SIC_3, CIC_3	0.969, 0.698, 0.665
3	3	χ_{PC}^b, χ_C^b	0.951, 0.631
4	4	SIC_1, O	0.922, 0.478
5	5	χ_{CH}^b	0.944
6	6	P_0	0.990
7	7	χ_{PC}^b	0.919
8	8	IC_3, IC, IC_2	0.909, 0.698, 0.537

^bNatural log of number of revertants per nanomole.

$$(48) \quad 41.572(SIC_4) + 2.636(\chi) + 3.728(\chi_{PC}^b) + 3.018(\chi_C^b)$$

$$(N = 95, r = 0.872, s = 0.98, F = 34.2)$$

Addition of the geometric parameters resulted in the following model:

$$\ln R = 15.785 + 3.883(\overline{IC}) - 1.374(O) - 14.152(SIC_1) + 2.878(\chi) +$$

$$(49) \quad 3.409(\chi_{PC}^b) + 4.625(\chi_C^b) - 7.867(P_0) - 0.0021(^3D W_H) + 0.0096(^3D W)$$

$$(N = 95, r = 0.893, s = 0.91, F = 37.2)$$

The predicted mutagenicity values for each of the aromatic amine chemicals from equations (48) and (49) are presented in Table 6.

4.6. DISCUSSION

The objectives of this chapter were to review the utility of TIs and 3D parameters in QSARs as well as to report recent results on the relative effectiveness of QSARs for estimating properties. A large number of QSAR models summarized here show that graph-theoretic invariants correlate reasonably well with physicochemical, biomedicinal, toxicological, and biochemical properties of diverse congeneric sets of molecules. It is also clear that TIs and substructures have found successful applications in the quantification of molecular similarity, selection of analogues, and molecular similarity-based estimation of properties. Of special interest is the fact that the molecular similarity method developed by Basak *et al.*² has been successfully used in the discovery of a novel class of human immunodeficiency virus reverse transcriptase (HIV-RT) inhibitors, showing the utility of such nonempirically based methods in practical drug discovery. Examples of

QSARs using 3D descriptors also demonstrate that geometrical parameters alone can predict properties of congeneric molecules quite satisfactorily.

In this context, it was of interest to compare the capabilities of TIs and 3D descriptors in QSAR analysis. To this end, we reported the QSAR studies developed on three different properties, *viz.*, normal boiling of 140 hydrocarbons, lipophilicity ($\log P$, octanol–water) of a diverse set of 254 chemicals, and mutagenicity ($\ln P$) of a group of 95 aromatic and heteroaromatic amines. Results of these QSARs show that TIs contain important structural information sufficient to develop useful predictive models for these properties. However, in the case of $\log P$, the addition of geometrical parameters, *viz.*, $3D_W$ and $3DW_H$, to the list of independent variables resulted in improved models in the sense that, while the TI-based QSARs had two outliers, the addition of geometrical variables gave well-behaved models including all of the hydrocarbons in the set. Also, estimate errors were significantly smaller for the regression equation using geometrical descriptors [equation (45) versus (44)]. This indicates that for $\log P$ of hydrocarbons, 3D or geometrical parameters encode some pertinent information relevant to $\log P$ which are not included in TIs. For $\log P$ and mutagenicity, however, the addition of geometrical descriptors resulted in only a slight improvement in the quality of the QSAR models over those derived from TIs only.

For $\log P$, we also used HB_1 , an algorithmically derived hydrogen bonding parameter, in addition to TIs, V_W , $3DW$, and $3DW_H$. This is because the magnitude of $\log P$ of a molecule is known to depend significantly on the strength of hydrogen-bonding ability of solutes with solvents.^{80,133} Our earlier studies on the correlation of $\log P$ using algorithmically derived parameters show that, HB_1 is an important parameter in predicting $\log P$ reported in this chapter provide evidence that the role of HB_1 , cannot be carried out by a combination of TIs and 3D parameters.

In many practical situations of drug design and risk assessment, one has to estimate physical/biomedical/toxicological properties of chemicals without access to any empirical data.³⁵ Similarity-based models^{1–3,28,29,47,49,50,66,111} and estimated values based on nonempirical structural parameters^{25,27,31,32,45,70} are two viable alternatives for deriving property values under such data-poor situations. The QSAR models reported here based on TIs and geometrical parameters may find applications in selecting analogues and in estimating properties of chemicals in such cases.

ACKNOWLEDGMENTS

This chapter is contribution No. 159 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this chapter was supported, in part, by grant F49620-94-1-0401 from the United States Air Force, Exxon Corporation, and the Structure–Activity Relationship Consortium (SARCON) of the Natural Resources Research Institute at the University of Minnesota.

REFERENCES

1. M. A. Johnson, S. C. Basak, and G. Maggiore, *Math. Comput. Modeling* **11**, 630 (1988).
2. S. C. Basak, V. R. Magnuson, G. J. Niemi, and R. R. Regal, *Discrete Appl. Math.* **19**, 17 (1988).
3. M. S. Lajiness, in: *Computational Chemical Graph Theory* (D. H. Rouvray, ed.), pp. 299–316. Nova, New York (1990).
4. R. E. Carhart, D. H. Smith, and R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **25**, 64 (1985).
5. S. C. Basak, C. Raychaudhury, A. B. Roy, and J. J. Ghosh, *Indian J. Pharmacol.* **13**, 112 (1981).
6. S. K. Ray, S. C. Basak, C. Raychaudhury, A. B. Roy, and J. J. Ghosh, *Arzneim. Forsch.* **32**, 322 (1982).
7. S. C. Basak, S. K. Ray, C. Raychaudhury, A. B. Roy, and J. J. Ghosh, *IRCS Med. Sci.* **10**, 145 (1982).
8. A. K. Samanta, S. K. Ray, S. C. Basak, and S. K. Bose, *Arzneim. Forsch.* **32**, 1515 (1982).
9. S. K. Ray, S. C. Basak, C. Raychaudhury, A. B. Roy, and J. J. Ghosh, *Indian J. Chem.* **20B**, 894 (1981).
10. S. C. Basak, D. P. Gieschen, V. R. Magnuson, and D. K. Harris, *IRCS Med. Sci.* **10**, 619 (1982).
11. S. K. Ray, S. C. Basak, C. Raychaudhury, A. B. Roy, and J. J. Ghosh, *Arzneim. Forsch.* **33**, 352 (1983).
12. S. C. Basak, D. P. Gieschen, D. K. Harris, and V. R. Magnuson, *J. Pharm.* **72**, 934 (1983).
13. S. C. Basak, D. K. Harris, and V. R. Magnuson, *J. Pharm.* **73**, 429 (1984).
14. A. B. Roy, C. Raychaudhury, S. K. Ray, S. C. Basak, and J. J. Ghosh, in: *Proceedings of the Fourth European Symposium on Chemical Structure–Biological Activity: Quantitative Approaches* (J. C. Deardorff, ed.), pp. 75–76. Elsevier, Amsterdam (1983).
15. S. K. Ray, S. Gupta, S. C. Basak, C. Raychaudhury, A. B. Roy, and J. J. Ghosh, *Indian J. Chem.* **24B**, 1149 (1985).
16. S. C. Basak, L. J. Monstrud, M. E. Rosen, C. M. Frane, and V. R. Magnuson, *Acta Pharm. Jugosl.* **36**, 81 (1986).
17. S. C. Basak, B. D. Gute, and L. R. Drewes, *Pharm. Res.* **13**, 775 (1996).
18. S. C. Basak, *Med. Sci. Res.* **15**, 605 (1987).
19. S. C. Basak, in: *Proceedings of the NATO Advanced Study Institute (ASI) on Pharmacokinetics*, Erice, Sicily, April 4–17, 1994. Plenum, New York.
20. R. Nilakantan, N. Bauman, and R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **31**, 527 (1991).
21. D. E. Needham, J. C. Wei, and P. G. Seybold, *J. Am. Chem. Soc.* **110**, 4186 (1988).
22. A. T. Balaban, *Chem. Phys. Lett.* **89**, 399 (1982).
23. A. T. Balaban, *Pure Appl. Chem.* **55**, 199 (1983).
24. A. T. Balaban, *MATCH* **21**, 115 (1986).
25. A. T. Balaban, S. C. Basak, T. Colburn, and G. D. Grunwald, *J. Chem. Inf. Comput. Sci.* **34**, 1118 (1994).
26. A. T. Balaban, *J. Chem. Inf. Comput. Sci.* **35**, 339 (1995).
27. S. C. Basak, G. J. Niemi, and G. D. Veith, in: *Computational Chemical Graph Theory* (D. H. Rouvray, ed.), p. 235. Nova, New York (1990).
28. S. C. Basak and G. D. Grunwald, *J. Chem. Inf. Comput. Sci.* **35**, 366 (1995).
29. S. C. Basak and G. D. Grunwald, *New J. Chem.* **19**, 231 (1995).
30. A. T. Balaban, S. Bertelsen, and S. C. Basak, *Math. Chem.* **30**, 55 (1994).
31. S. C. Basak, G. J. Niemi, and G. D. Veith, *J. Math. Chem.* **4**, 185 (1990).
32. S. C. Basak, G. J. Niemi, and G. D. Veith, *Math. Comput. Modelling* **14**, 511 (1990).
33. S. C. Basak, G. J. Niemi, and G. D. Veith, *J. Math. Chem.* **7**, 243 (1991).
34. D. Bonchev and N. Trinajstić, *J. Chem. Phys.* **67**, 4517 (1977).
35. M. Randit, *J. Am. Chem. Soc.* **97**, 6609 (1975).
36. C. Raychaudhury, S. K. Ray, J. J. Ghosh, A. B. Roy, and S. C. Basak, *J. Comput. Chem.* **5**, 581 (1984).
37. D. H. Rouvray and R. B. Pandey, *J. Chem. Phys.* **85**, 2288 (1986).
38. D. H. Rouvray, *New Sci. May*, 35 (1993).
39. K. Balasubramanian, *SAR QSAR Environ. Res.* **2**, 59 (1994).
40. M. Randić, *Int. J. Quantum Chem. Quant. Biol. Symp.* **11**, 137 (1984).
41. S. C. Basak, D. P. Gieschen, and V. R. Magnuson, *Environ. Toxicol. Chem.* **3**, 191 (1984).

42. S. C. Basak, C. M. Frane, M. E. Rosen, and V. R. Magnuson, *IRCS Med. Sci.* **14**, 848 (1986).

43. S. C. Basak, *Med. Sci. Res.* **16**, 281 (1988).

44. G. J. Niemi, S. C. Basak, and G. D. Veith, in: *Envirotech Vienna: Proceedings of the First Conference of the International Society of Environmental Protection* (K. Zirm and J. Mayer, eds.), pp. 57–68. W. B. Druck GmbH and Co., Reiden, Austria (1989).

45. G. J. Niemi, S. C. Basak, G. D. Veith, and G. D. Grunwald, *Toxicol. Chem.* **11**, 893 (1992).

46. S. C. Basak, S. Bertelsen, and G. D. Grunwald, *Toxicol. Lett.* **79**, 239 (1995).

47. S. C. Basak and G. D. Grunwald, *SAR QSAR Environ. Res.* **2**, 289 (1994).

48. S. C. Basak and G. D. Grunwald, in: *Proceeding of the XVI International Cancer Congress* (R. S. Rao, M. G. Deo, and L. D. Sanghvi, eds.), pp. 413–416. Monduzzi, Bologna, Italy (1995).

49. S. C. Basak and G. D. Grunwald, *SAR QSAR Environ. Res.* **3**, 265 (1995).

50. S. C. Basak and G. D. Grunwald, *Chemosphere* **31**, 2529 (1995).

51. T. Colburn, D. Axel, and S. C. Basak, *Mutat. Res.* (in preparation).

52. S. C. Basak, in: *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology* (W. Karcher and J. Devillers, eds.), pp. 83–103. Kluwer Academic, Dordrecht (1990).

53. S. C. Basak, G. D. Grunwald, G. E. Host, G. J. Niemi, and S. Bradbury, *Environ. Toxicol. Chem.* (in preparation).

54. C. Hansch, *Adv. Pharmacol. Chemother.* **13**, 45 (1975).

55. C. M. Auer, J. V. Nabholz, and K. P. Baetke, *Environ. Health Perspect.* **87**, 183 (1990).

56. J. C. Arcos, *Environ. Sci. Technol.* **21**, 743 (1987).

57. U. Burkert and N. L. Allinger, *Molecular Mechanics, ACS Monograph 177*. American Chemical Society, Washington, DC (1982).

58. W. G. Richards, *Quantum Pharmacology*, Butterworths, London (1977).

59. A. Verloof, W. Hoogenstraaten, and J. Tipker, in: *Drug Design*, Vol. VII (E. J. Ariens, ed.), pp. 165–207. Academic Press, New York (1976).

60. M. J. Kamlet, R. M. Doherty, G. D. Veith, R. W. Taft, and M. H. Abraham, *Environ. Sci. Technol.* **20**, 690 (1986).

61. Moriguchi and Y. Kanada, *Chem. Pharm. Bull.* **25**, 926 (1977).

62. M. Bunge, *Methods, Models and Mater.*, Reidel, Dordrecht (1973).

63. F. Harary, *Graph Theory*, Addison-Wesley, Reading, Massachusetts (1969).

64. N. Trinajstić, *Chemical Graph Theory*, CRC Press, Boca Raton, Florida (1983).

65. I. S. Dmitriev, *Molecules Without Chemical Bonds*, Mir Publishers, Moscow (1981).

66. M. Randić, in: *Concepts and Applications of Molecular Similarity* (M. A. Johnson and G. M. Maggiora, eds.), pp. 77–145. John Wiley & Sons, New York (1990).

67. C. Raychaudhury, S. K. Ray, J. J. Ghosh, A. B. Roy, and S. C. Basak, *J. Comput. Chem.* **5**, 581 (1984).

68. A. Sablić and N. Trinajstić, *Acta Pharm. Jugosl.* **31**, 189 (1981).

69. O. Mekenyán, S. Dimitrov, and D. Bonchev, *Eur. Polym. J.* **19**, 1185 (1983).

70. A. T. Balaban, N. Joshi, L. B. Kier, and L. H. Hall, *J. Chem. Inf. Comput. Sci.* **32**, 233 (1992).

71. M. V. Duudea, O. Minailuc, and A. T. Balaban, *J. Comput. Chem.* **12**, 527 (1991).

72. A. T. Balaban, *Theor. Chim. Acta* **53**, 355 (1979).

73. P. A. Filip, T. S. Balaban, and A. T. Balaban, *J. Math. Chem.* **1**, 61 (1987).

74. A. T. Balaban and V. Feroiu, *Rep. Mol. Theory J.* **1**, 133 (1990).

75. E. J. Kupchik, *Quanti. Struct. Act. Relat.* **7**, 57 (1988).

76. L. Pogliani, *J. Phys. Chem.* **97**, 6731 (1993).

77. L. B. Kier and L. H. Hall, *Molecular Connectivity in Structure-Activity Analysis*, Research Studies Press, New York (1986).

78. S. C. Basak and G. D. Grunwald, *Math. Modelling Sci. Comput.* **2**, 735 (1993).

79. D. Bonchev and N. Trinajstić, *Int. J. Quantum Chem.* **12**, 293 (1978).

80. S. C. Basak, B. D. Gute, and S. Ghatak, *J. Chem. Inf. Comput. Sci.* (submitted for publication).

81. S. C. Basak and V. R. Magnuson, *Arzneim. Forsch.* **33**, 501 (1983).

82. J. V. Soderman, *CRC Handbook of Identified Carcinogens and Noncarcinogens: Carcinogenicity—Mutagenicity Database*, Vol. I, CRC Press, Boca Raton, Florida (1982).

83. S. C. Basak, C. M. Frane, M. E. Rosen, and V. R. Magnuson, *Med. Sci. Res.* **15**, 887 (1987).

84. V. R. Magnuson, D. K. Harris, and S. C. Basak, in: *Studies in Physical and Theoretical Chemistry* (R. B. King, ed.), pp. 178–191. Elsevier, Amsterdam (1983).

85. F. C. Smeeks and P. C. Jurs, *Theor. Chim. Acta* **233**, 111 (1990).

86. Y. Gao and H. Hosoya, *Bull. Chem. Soc. Jpn.* **61**, 3093 (1988).

87. H. Wiener, *J. Am. Chem. Soc.* **69**, 17 (1947).

88. J. R. Platt, *J. Chem. Phys.* **15**, 419 (1947).

89. L. H. Hall and L. B. Kier, *Tetrahedron* **33**, 1953 (1977).

90. L. Pogliani, *J. Phys. Chem.* **99**, 925 (1995).

91. W. J. Boecklen and G. J. Niemi, *SAR QSAR Environ. Res.* **2**, 79 (1994).

92. A. T. Balaban and C. Catana, *SAR QSAR Environ. Res.* **2**, 1 (1994).

93. S. P. Gupta and P. Singh, *Bull. Chem. Soc. Jpn.* **52**, 2745 (1979).

94. M. Randić, *New J. Chem.* **15**, 517 (1991).

95. R. H. Rohrbaugh and P. C. Jurs, *Anal. Chem.* **60**, 2249 (1988).

96. D. K. Pal, S. K. Purkyaastha, C. Sengupta, and A. U. De, *Indian J. Chem.* **31**, 109 (1992).

97. D. K. Pal, S. K. Purkyaastha, C. Sengupta, and A. U. De, *Indian J. Chem.* **33**, 247 (1989).

98. D. H. Rouvray and W. Tatong, *Int. J. Environ. Stud.* **33**, 247 (1989).

99. G. J. Niemi, R. R. Regal, and G. D. Veith, *ACS Symp. Ser.* **292**, 148 (1985).

100. I. Lukovits, *J. Chem. Soc. Perkin Trans. 2*, 1667 (1988).

101. G. J. Niemi, G. D. Veith, R. R. Regal, and D. D. Vaishnav, *Environ. Toxicol. Chem.* **6**, 515 (1987).

102. V. K. Gombar and K. Einstlein, in: *Applied Multivariate Analysis in SAR and Environmental Studies* (J. Devillers and W. Karcher, eds.), pp. 377–414. Kluwer Academic, Dordrecht (1991).

103. B. W. Blake, K. Einstlein, V. K. Gombar, and H. H. Borgstedt, *Mutat. Res.* **241**, 261 (1990).

104. T. Okuyama, Y. Miyashita, S. Kanaya, H. Katsumi, S. Sasaki, and M. Randić, *J. Comput. Chem.* **9**, 636 (1988).

105. C. L. Wilkins and M. Randić, *Theor. Chim. Acta* **58**, 45 (1980).

106. M. Randić and N. Trinajstić, *MATCH* **13**, 271 (1982).

107. M. Randić, in: *Molecular Basis of Cancer. Part A: Macromolecular Structure, Carcinogens, and Oncogenes* (R. Rein, ed.), pp. 309–318. Alan R. Liss, New York (1985).

108. S. C. Basak, S. Bertelsen, and G. D. Grunwald, *J. Chem. Inf. Comput. Sci.* **34**, 270 (1994).

109. A. Leo and D. Weininger, *CLOG Version 3.2 User Reference Manual*, Medicinal Chemistry Project, Pomona College, Claremont, California (1984).

110. P. Willett, *J. Chem. Inf. Comput. Sci.* **23**, 22 (1983).

111. P. Willett and V. Winterman, *Quant. Struct. Act. Relat.* **5**, 18 (1986).

112. P. Willett, in: *Concepts and Applications of Molecular Similarity* (M. A. Johnson and G. M. Maggiora, eds.), pp. 43–63. John Wiley & Sons, New York (1990).

113. G. M. Downs and P. Willett, in: *Applied Multivariate Analysis in SAR and Environmental Studies* (J. Devillers and W. Karcher, eds.), pp. 247–279. Kluwer Academic, Dordrecht (1991).

114. P. A. Bath, A. R. Andrew, and P. Willett, *J. Chem. Inf. Comput. Sci.* **34**, 141 (1994).

115. R. D. Brown, G. Jones, and P. Willett, *J. Chem. Inf. Comput. Sci.* **24**, 63 (1984).

116. S. C. Basak, B. D. Gute, and G. D. Grunwald, *Croat. Chem. Acta* **69** (1996) (in press).

117. J. E. Amoore, *Nature* **214**, 1095 (1967).

118. M. Charlton, *Top. Curr. Chem.* **14**, 107 (1983).

119. B. Bogdanov, S. Nikolic, and N. Trinajstić, *J. Math. Chem.* **3**, 299 (1989).

120. R. P. Bhatnagar, P. Singh, and S. P. Gupta, *Indian J. Chem.* **19B**, 780 (1980).

121. R. D. Cramer III, D. E. Patterson, and J. D. Bunce, *J. Am. Chem. Soc.* **110**, 5959 (1988).

122. O. Mekenyán, D. Bonchev, and N. Trinajstić, *Int. J. Quantum Chem.* **18**, 369 (1980).

123. W. Karcher, *Spectral Atlas of Polycyclic Aromatic Hydrocarbons*, Vol. 2, pp. 16–19. Kluwer Academic, Dordrecht (1988).

124. A. K. Debnath, G. Debnath, A. J. Shusterman, and C. Hansch, *Environ. Mol. Mutagen.* **19**, 37 (1992).

125. S. C. Basak, D. K. Harriss, and V. R. Magnuson, POLLY 2.3, copyright by the University of Minnesota (1988).

126. C. E. Shannon, *Bell Syst. Tech. J.* **27**, 379 (1948).

127. R. Sarkar, A. B. Roy, and R. K. Sarkar, *Math. Biosci.* **39**, 379 (1978).

128. A. B. Roy, S. C. Basak, D. K. Harriss, and V. R. Magnuson, in: *Mathematical Modelling in Science and Technology* (K. J. R. Ayula, R. E. Kalman, A. I. Liapis, and E. Y. Rodin, eds.), pp. 745-750, Pergamon Press, Elmsford, New York (1984).

129. S. C. Basak, A. B. Roy, and J. J. Ghosh, in: *Proceedings of the Second International Conference on Mathematical Modelling*, Vol. II (X. J. R. Avila, R. Bellman, Y. L. Luke, and A. K. Rigler, eds.), pp. 851-856, University of Missouri, Rolla (1980).

130. Tripos Associates, Inc., SYBYL Version 6.1, Tripos Associates, Inc., St. Louis, Missouri (1994).

131. Tripos Associates, Inc., CONCORD Version 3.01, Tripos Associates, Inc., St. Louis, Missouri (1993).

132. S. C. Basak, H-BOND: A Program for Calculating Hydrogen Bonding Parameter, University of Minnesota, Duluth (1988).

133. Y.-C. Ou, Y. Ouyang, and E. J. Lien, *J. Mol. Sci.* **4**, 89 (1986).

134. SAS Institute, Inc., SAS/STAT User's Guide, Release 6.03 Edition, SAS Institute, Inc., Cary, North Carolina (1988).

135. R. R. Hocking, *BiometRICS* **32**, 1 (1976).

136. S. Weisberg, *Applied Linear Regression*, John Wiley & Sons, New York (1980).

CHARACTERIZATION OF THE MOLECULAR SIMILARITY
OF CHEMICALS USING TOPOLOGICAL INVARIANTS

Subhash C. Basak*

Brian D. Gute

and

Gregory D. Grunwald

Center for Water and the Environment
Natural Resources Research Institute
University of Minnesota, Duluth
5013 Miller Trunk Highway
Duluth, MN 55811, USA

Advances in Molecular Similarity, JAI Press, submitted, 1997.

* To whom all correspondence should be addressed.

Abstract

Three similarity spaces were used in the selection of analogs and K -nearest neighbor (KNN) based estimation of normal boiling points for a diverse set of 2926 chemicals. The similarity spaces consisted of principal components (PCs) derived from: 1) 40 topostructural indices, 2) 61 topochemical parameters and 3) the full set 101 topostructural and topochemical indices. The three methods selected sets of analogs with a substantial number of structurally analogous molecules. For the KNN method of property estimation, the similarity space which used the full set of indices was superior to either of the subsets (topostructural or topochemical). For all three methods, $K = 6-10$ gave the best estimated values for boiling point.

1. Introduction

Interest in quantifying the similarity of molecules using computational methods has increased [1-8]. In particular, a recent trend in the characterization of similarity/dissimilarity of chemicals makes use of graph invariants. Molecular structures can be represented by planar graphs, $G = [V, E]$, where the nonempty set V represents the set of atoms and the set E generally represents covalent bonds [9]. These graphs can be used to adequately represent the pattern of connectedness of atoms within a molecule. Graph invariants, values derived from planar graphs, are graph theoretic properties which are identical for isomorphic graphs. A numerical graph invariant or topological index maps a chemical structure into the set of real numbers.

Various graph invariants have been used in ordering and partial ordering of sets of molecules [1, 4-8]. Various topological indices (TIs) and principal components (PCs) derived from TIs have been used in quantifying the similarity/dissimilarity of molecules and in the similarity based estimation of physical and toxicological properties [4, 5, 10-17]. Such TIs include those derived from simple planar graphs which contain adjacency and distance information for vertices. These TIs could be considered topostructural indices. Other TIs, which are derived from weighted chemical graphs, could be called topochemical indices because they contain explicit information regarding the chemical nature of the atoms (vertices) and bonds (edges) in the molecular structure, in addition to quantifying the adjacency and distance relationships within the graph.

Our earlier studies made use of a combination of topostructural and topochemical indices to select analogs of chemicals and estimate properties of

molecules in large and diverse databases using the K-nearest neighbor (KNN) method.

In this paper we have carried out a comparative analysis of similarity based analog selection and KNN based estimation of normal boiling point using : a) a set of 40 topostructural indices, b) a group of 61 topochemical indices, and c) the combined set of 101 indices.

2. Methods

2.1 DATABASE

The normal boiling point database consisted of 2926 compounds taken from the U.S. EPA ASTER [18] system. This data comprised a set for which chemical structures and normal boiling values were available, and for which it was possible to compute all 101 TIs.

2.2 CALCULATION OF INDICES

The TIs calculated for this study are listed in table 1 and include Wiener number [19], molecular connectivity indices as calculated by Randić [20] and Kier and Hall [21], frequency of path lengths of varying size, information theoretic indices defined on distance matrices of graphs using the methods of Bonchev and Trinajstić [22] as well as those of Raychaudhury et al. [23], parameters defined on the neighborhood complexity of vertices in hydrogen-filled molecular graphs [24-26], and Balaban's *J* indices [27-29]. The majority of the TIs were calculated using POLLY 2.3 [30]. The *J* indices were calculated using software developed by the authors.

The Wiener index (W), the first topological index reported in the chemical literature [19], may be calculated from the distance matrix $D(G)$ of a hydrogen-suppressed chemical graph G as the sum of the entries in the upper triangular distance submatrix. The distance matrix $D(G)$ of a nondirected graph G with n vertices is a symmetric $n \times n$ matrix (d_{ij}) , where d_{ij} is equal to the distance between vertices v_i and v_j in G . Each diagonal element d_{ii} of $D(G)$ is zero. We give below the distance matrix $D(G_1)$ of the unlabeled hydrogen-suppressed graph G_1 of *n*-propanol (figure 1):

$$D(G_1) = \begin{matrix} & \begin{matrix} (1) & (2) & (3) & (4) \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \left[\begin{matrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{matrix} \right] \end{matrix}$$

W is calculated as:

$$W = \frac{1}{2} \sum_{i,j} d_{ij} = \sum_h h \cdot g_h \quad (1)$$

where g_h is the number of unordered pairs of vertices whose distance is h . Thus for $D(G_1)$, W has a value of ten.

[Insert Fig. 1 here]

Randić's connectivity index [20], and higher-order connectivity path, cluster, path-cluster and chain types of simple, bond and valence connectivity parameters were

calculated using the method of Kier and Hall [21]. The generalized form of the simple path connectivity index is as follows:

$$^h\chi = \sum_{\text{paths}} (v_i v_j \dots v_{h+1})^{-h} \quad (2)$$

where v_i, v_j, \dots, v_{h+1} are the degrees of the vertices in the path of length h . The path length parameters (P_h), number of paths of length h ($h = 0, 1, \dots, 10$) in the hydrogen-suppressed graph, are calculated using standard algorithms.

Information-theoretic topological indices are calculated by the application of information theory on chemical graphs. An appropriate set A of n elements is derived from a molecular graph G depending upon certain structural characteristics. On the basis of an equivalence relation defined on A , the set A is partitioned into disjoint subsets A_i of order n_i ($i = 1, 2, \dots, h$; $\sum_i n_i = n$). A probability distribution is then assigned to the set of equivalence classes:

$$A_1, A_2, \dots, A_h$$

$$p_1, p_2, \dots, p_h$$

where $p_i = n_i / n$ is the probability that a randomly selected element of A will occur in the i^{th} subset.

The mean information content of an element of A is defined by Shannon's relation [31]:

$$IC = - \sum_{i=1}^h p_i \log_2 p_i \quad (3)$$

The logarithm is taken at base 2 for measuring the information content in bits. The total information content of the set A is then $n \times IC$.

To account for the chemical nature of vertices as well as their bonding pattern, Sarkar *et al.* [32] calculated information content of chemical graphs on the basis of an equivalence relation where two atoms of the same element are considered equivalent if they possess an identical first-order topological neighborhood. Since properties of atoms or reaction centers are often modulated by stereo-electronic characteristics of distant neighbors, *i.e.*, neighbors of neighbors, it was deemed essential to extend this approach to account for higher-order neighbors of vertices. This can be accomplished by defining open spheres for all vertices of a chemical graph. If r is any non-negative real number and v is a vertex of the graph G , then the open sphere $S(v, r)$ is defined as the set consisting of all vertices v_i in G such that $d(v, v_i) < r$. Therefore, $S(v, 0) = \phi$, $S(v, r) = v$ for $0 < r < 1$, and $S(v, r)$ is the set consisting of v and all vertices v_i of G situated at unit distance from v , if $1 < r < 2$.

One can construct such open spheres for higher integral values of r . For a particular value of r , the collection of all such open spheres $S(v, r)$, where v runs over the whole vertex set V , forms a neighborhood system of the vertices of G . A suitably defined equivalence relation can then partition V into disjoint subsets consisting of vertices which are topologically equivalent for r^{th} order neighborhood. Such an

approach has been developed and the information-theoretic indices calculated based on this idea are called indices of neighborhood symmetry [26].

In this method, chemicals are symbolized by weighted linear graphs. Two vertices u_o and v_o of a molecular graph are said to be equivalent with respect to r^{th} order neighborhood if and only if corresponding to each path u_o, u_1, \dots, u_r of length r , there is a distinct path v_o, v_1, \dots, v_r of the same length such that the paths have similar edge weights, and both u_o and v_o are connected to the same number and type of atoms up to the r^{th} order bonded neighbors. The detailed equivalence relation has been described in earlier studies [26, 33].

Once partitioning of the vertex set for a particular order of neighborhood is completed, IC_r is calculated by eq. (2). Basak *et al.* defined another information-theoretic measure, structural information content (SIC_r), which is calculated as:

$$SIC_r = IC_r / \log_2 n \quad (4)$$

where IC_r is calculated from eq. (2) and n is the total number of vertices of the graph [24].

Another information-theoretic invariant, complementary information content (CIC_r), is defined as:

$$CIC_r = \log_2 n - IC_r \quad (5)$$

C/C_r represents the difference between maximum possible complexity of a graph (where each vertex belongs to a separate equivalence class) and the realized topological information of a chemical species as defined by IC_r [25].

In figure 2, the calculation of IC_2 , SIC_2 and C/C_2 is demonstrated for the labeled hydrogen-filled graph (G_2) of *n*-propanol.

[Insert Fig. 2 here]

The information-theoretic index on graph distance, I_D^W is calculated from the distance matrix $D(G)$ of a chemical graph G as follows [22]:

$$I_D^W = W \log_2 W - \sum_h g_h \cdot h \log_2 h \quad (6)$$

The mean information index, \bar{I}_D^W , is found by dividing the information index I_D^W by W . The information theoretic parameters defined on the distance matrix, H^D and H^V , were calculated by the method of Raychaudhury *et al* [23].

Balaban defined a series of indices based upon distance sums within the distance matrix for a chemical graph which he designated as J indices [27-29]. These indices are highly discriminating with low degeneracy. Unlike W , the J indices range of values are independent of molecular size. The general form of the J index calculation is as follows:

$$J = q(\mu + 1)^{-1} \sum_{i,j, \text{edges}} (s_i s_j)^{-\gamma} \quad (7)$$

where the cyclomatic number μ (or number of rings in the graph) is $\mu = q-n+1$, with q edges and n vertices and s_i is the sum of the distances of atom i to all other atoms and s_j is the sum of the distances of atom j to all other atoms [27]. Variants were proposed by Balaban for incorporating information on bond type, relative electronegativities, and relative covalent radii [28-29].

2.3 CLASSIFICATION OF THE INDICES

The set of 101 TIs was partitioned into two distinct subsets: topostructural indices and topochemical indices. Topostructural indices encode information about the adjacency and distances of atoms (vertices) in molecular structures (graphs) irrespective of atom type or factors such as hybridization states and number of core/valence electrons in individual atoms. Topochemical indices quantify information regarding specific chemical properties of the atoms comprising a molecule as well as the topology (connectivity of atoms). Topochemical indices are derived from weighted molecular graphs where each vertex (atom) is properly weighted with selected chemical/physical properties. These subsets are shown in table 1.

2.4 STATISTICAL METHODS AND COMPUTATION OF SIMILARITY

Data Reduction

Initially, all TIs were transformed by the natural logarithm of the index plus one. This was done since the scale of some TIs may be several orders of magnitude greater than other TIs.

A principal component analysis (PCA) was used on the transformed indices to minimize intercorrelation of indices. The PCA analysis was accomplished using the SAS procedure PRINCOMP [34]. The PCA produces linear combinations of the TIs, called principal components (PCs) which are derived from the correlation matrix. The first PC has the largest variance, or eigenvalue, of the linear combination of TIs. Each subsequent PC explains the maximal index variance orthogonal to the previous PCs, eliminating any redundancies which could occur within the set of TIs. The maximum number of PCs generated is equal to the number of TIs available. For the purposes of this study, only PCs with eigenvalues greater than one were retained. A more detailed explanation of this approach has been provided in a previous study by Basak *et al* [4]. These PCs were subsequently used in determining similarity scores as described below.

Similarity Measures

Intermolecular similarity was measured by the Euclidean distance (ED) within an *n*-dimensional space. This *n*-dimensional space consisted of orthogonal variables (PCs) derived from the TIS as described above. ED between the molecules *i* and *j* is defined as:

$$ED_{ij} = \left[\sum_{k=1}^n (D_{ik} - D_{jk})^2 \right]^{1/2} \quad (8)$$

where *n* equals the number of dimensions or PCs retained from the PCA. D_{ik} and D_{jk} are the data values of the *k*th dimension for chemicals *i* and *j*, respectively.

K-Nearest Neighbor Selection and Property Estimation

Following the quantification of intermolecular similarity of the 2926 chemicals, the *K*-nearest neighbors ($K = 1-10, 15, 20, 25$) were determined on the basis of ED. This procedure can be used to select structural analogs (neighbors) of a probe compound or the neighbors can be used in property estimation. In estimating the normal boiling point of the probe compound, the mean observed normal boiling point of the *K*-nearest neighbors was used as the estimate and the standard error (*s*) of the estimate was used to assess the efficacy of the set of indices.

3. Results

3.1 PRINCIPAL COMPONENT ANALYSIS

From the PCA of the 40 topostructural indices, seven PCs with eigenvalues greater than one were retained. These seven PCs explained, cumulatively, 90.8% of the total variance within the TI data. Table 2 lists the eigenvalues of the seven PCs, the proportion of variance explained by each PC, the cumulative variance explained, and the three TIs most correlated with each individual PC.

The PCA of the 61 topochemical indices resulted in the selection of ten PCs, all having eigenvalues greater than one. The ten PCs explain a total of 92.1% of the variance within the TI data. Table 3 presents a summary of the information regarding these ten PCs.

Twelve PCs were retained from the PCA of the full set of 101 TIs. Each of these

PCs had an eigenvalue greater than one and, cumulatively, they explained 92.8% of the variance within the full set of TIs. These PCs are summarized in table 4.

3.2 ANALOG SELECTION

Figure 3 shows an example of analog selection using PCs to derive a Euclidean distance space. The first five analogs (neighbors) for the probe compound, 3-methyl-4-chlorophenol, are presented for each of the three similarity spaces. The analogs selected by the topostructural model show a repetition of the same skeletal structure, ignoring substituents, throughout the first five analogs. In the topochemical model and the full set model some variability in the skeletal structure arises (chemical analogs 2 & 5, full set analog 4). Also of interest is the repetition of chemicals between the sets of analogs. While the ordering varies between the methods, the topostructural and topochemical models select two identical structures, the topostructural and the full set have three analogs in common, and the topochemical and full set select four of the same analogs. 2-chloro-5-methylphenol appears in all three sets, while there are only three unique compounds (topostructural analogs 4 & 5, topochemical analog 5).

[Insert Fig. 3]

3.3 K-NEAREST NEIGHBOR PROPERTY ESTIMATION

Figure 4 presents the correlation (r) and the standard error (s) of the prediction of the normal boiling points for the 2926 chemicals for the three groups of indices over the full range of K values examined ($K = 1-10, 15, 20, 25$). Table 5 shows the best normal

boiling point model for each set of indices. The best boiling point estimates for all three sets were for K in the range of 6 to 10. The full set of indices gave the best result, however, there was only a small difference between models.

[Insert Fig. 4]

4. Discussion

The purpose of this paper was to study the relative effectiveness of three similarity spaces derived from graph invariants in the selection of structural analogs and in the KNN based estimation of properties. The similarity spaces were created using a principal component analysis of calculated graph invariants. Tables 2-4 summarize the results of the PCA of the three sets of indices. The first PC is always correlated with indices which quantify molecular size. In the case of the topostructural indices, the second PC is most correlated with branching indices. In the case of PCs derived from either topochemical or the full set of topostructural and topochemical parameters, the first PC was strongly correlated with molecular size, while the second PC was highly associated with the molecular complexity indices. These results are in line with our earlier studies on different sets of chemicals [4, 5, 11, 35, 36].

All three spaces were used in the selection of five analogs of a particular structure (Figure 3). Perusal of the three sets of structures show that there is a substantial degree of similarity among the three groups of five chemicals selected. It is interesting to note that all five nearest neighbors of the probe selected by the topostructural method had isomorphic skeletal graphs when hydrogen atoms are

suppressed. For the two similarity spaces created by topochemical indices alone and the combined set of topostructural and topochemical indices, four of the five selected neighbors are common (Figure 3) although the ordering of the molecules is different. This shows that these two similarity methods are not intrinsically very different. Our earlier results showed that analogs selected by similarity methods derived from experimental physical properties, atom pairs and topological indices select very similar sets of analogs [10].

In the case of KNN based estimation of boiling points of chemicals from their analogs, K was varied from 1 to 25. The best estimated value was obtained in the range of $K = 6-10$. This is in line with our earlier studies with different properties [11, 12].

In conclusion, the three similarity spaces derived in this paper have reasonable power for selecting analogous molecules from a very diverse database of chemicals. The KNN based estimation shows that selected analogs can be used for the estimation of boiling points of diverse chemicals if more accurate methods are not available.

5. Acknowledgments

This is contribution number XXX from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported in part by grants F49620-94-1-0401 and F49620-96-1-0330 from the United States Air Force, a grant from Exxon Corporation and the Structure-Activity Relationship Consortium (SARCON) of the Natural Resources Research Institute of the University of Minnesota.

6. References

- [1] M.A. Johnson and G.M. Maggiora, eds., *Concepts and applications of molecular similarity* (Wiley, 1990).
- [2] R. Carbó, L. Leyda and M. Arnau, *Int. J. Quantum Chem.* 17(1980) 1185.
- [3] P.E. Bowen-Jenkins, D.L. Cooper and G. Richards, *J. Phys. Chem.* 89 (1985) 2195.
- [4] S.C. Basak, V.R. Magnuson, G.J. Niemi and R.R. Regal, *Discrete Appl. Math.* 19 (1988) 17.
- [5] S.C. Basak, S. Bertelsen and G. Grunwald, *J. Chem. Inf. Comput. Sci.* 34 (1994) 270.
- [6] G. Rum and W.C. Herndon, *J. Am. Chem. Soc.* 113 (1991) 9055.
- [7] P. Willett and V. Winterman, *Quant. Struct.-Act. Relat.* 5 (1986) 18.
- [8] C.L. Wilkins and M. Randić, *Theoret. Chim. Acta (Berl.)* 58 (1980) 45.
- [9] N. Trinajstić, *Chemical Graph Theory Vols. I & II* (CRC Press, Boca Raton, Florida, 1983).
- [10] S.C. Basak and G.D. Grunwald, *Mathl. Modelling Sci. Comput.*, in press.
- [11] S.C. Basak and G.D. Grunwald, *SAR QSAR Environ. Res.* 2 (1994) 289.
- [12] S.C. Basak and G.D. Grunwald, *New J. Chem.* 19 (1995) 231.
- [13] S.C. Basak and G.D. Grunwald, *J. Chem. Inf. Comput. Sci.* 35 (1995) 366.
- [14] S.C. Basak and G.D. Grunwald, *SAR QSAR Environ. Res.* 3 (1995) 265.
- [15] S.C. Basak and G.D. Grunwald, *Chemosphere* 31 (1995) 2529.
- [16] S.C. Basak, B.D. Gute and G.D. Grunwald, *Croat. Chim. Acta*, 69 (1996) 1159.
- [17] M.S. Lajiness, in: *Computational Chemical Graph Theory*, ed. D.H. Rouvray (Nova Science Publishers, New York, 1990) p. 300.
- [18] C.L. Russom, *Assessment Tools for the Evaluation of Risk (Aster) v. 1.0* (U.S. Environmental Protection Agency, 1992).

- [19] H. Wiener, *J. Am. Chem. Soc.* 69 (1947) 17.
- [20] M. Randić, *J. Am. Chem. Soc.* 97 (1975) 6609.
- [21] L.B. Kier and L.H. Hall, *Molecular Connectivity in Structure-Activity Analysis* (Research Studies Press, Hertfordshire, U.K., 1986).
- [22] D. Bonchev and N. Trinajstić, *J. Chem. Phys.* 67 (1977) 4517.
- [23] C. Raychaudhury, S.K. Ray, J.J. Ghosh, A.B. Roy and S.C. Basak, *J. Comput. Chem.* 5 (1984) 581.
- [24] S.C. Basak, A.B. Roy and J.J. Ghosh, in: *Proceedings of the Second International Conference on Mathematical Modelling*, eds. X.J.R. Avula, R. Bellman, Y.L. Luke and A.K. Rigler (University of Missouri - Rolla, 1980) p. 851.
- [25] S.C. Basak and V.R. Magnuson, *Arzneim.-Forsch. Drug Res.* 33 (1983) 501.
- [26] A.B. Roy, S.C. Basak, D.K. Harriss and V.R. Magnuson, in: *Mathematical Modelling in Science and Technology*, eds. X.J.R. Avula, R.E. Kalman, A.I. Liapis and E.Y. Rodin (Pergamon Press, New York, 1984) p. 745.
- [27] A.T. Balaban, *Chem. Phys. Lett.* 89 (1982) 399.
- [28] A.T. Balaban, *Pure and Appl. Chem.* 55 (1983) 199.
- [29] A.T. Balaban, *Math. Chem. (MATCH)* 21 (1985) 115.
- [30] S.C. Basak, D.K. Harriss and V.R. Magnuson, *POLLY v. 2.3* (Copyright of the University of Minnesota, 1988).
- [31] C.E. Shannon, *Bell Syst. Tech. J.* 27 (1948) 379.
- [32] R. Sarkar, A.B. Roy and R.K. Sarkar, *Math. Biosci.* 39 (1978) 299.
- [33] V.R. Magnuson, D.K. Harriss and S.C. Basak, in: *Studies in Physical and Theoretical Chemistry*, ed. R.B. King (Elsevier, Amsterdam, 1983) p. 178.
- [34] SAS Institute Inc, in: *SAS/STAT User's Guide, Release 6.03 Edition* (SAS Institute Inc., Cary, NC, 1988) p. 751.
- [35] S.C. Basak, G.J. Niemi and G.D. Veith, *J. Math. Chem.*, 7 (1991) 243.
- [36] S.C. Basak, V.R. Magnuson, G.J. Niemi, R.R. Regal and G.D. Veith, *Mathl. Modelling* 8 (1987) 300.

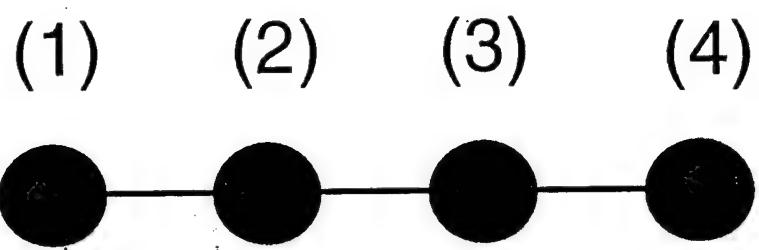
Figure Legend

Figure 1 The unlabeled hydrogen-suppressed graph (G_1) of *n*-propanol.

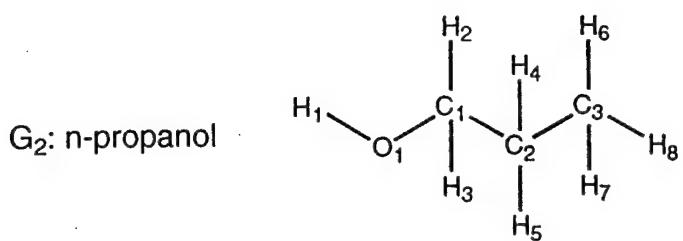
Figure 2 Calculation of the indices IC_2 , SC_2 , and CC_2 for the hydrogen-filled, labeled graph (G_2) of *n*-propanol.

Figure 3 The five analogs selected for the probe 3-methyl-4-chlorophenol using three molecular similarity spaces: topostructural, topochemical, and all indices. The numbers under the structures indicate the ranking of the analogs and the Euclidean distance to the probe.

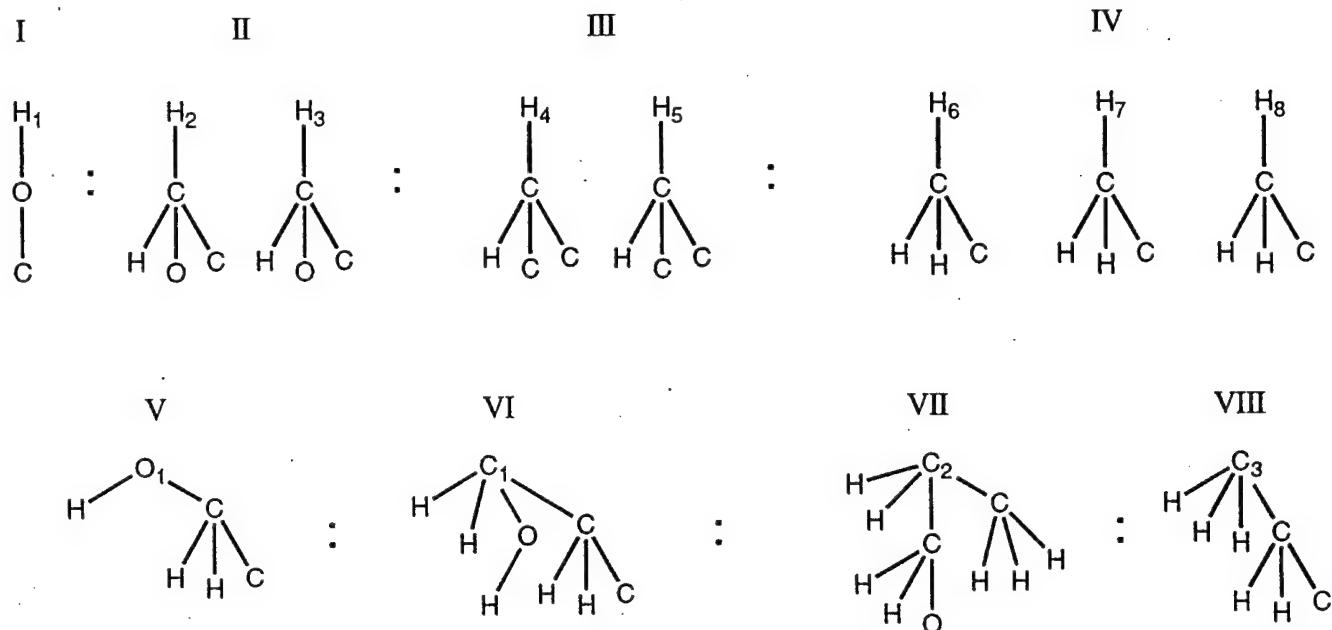
Figure 4 Pattern of: (a) correlation (r) and (b) standard error (s) of the estimates according to the K -nearest neighbor selection for 2926 normal boiling points using three molecular similarity spaces.



G_1



Second order neighbors:



Subsets:

I	II	III	IV	V	VI	VII	VIII
(H ₁)	(H ₂ -H ₃)	(H ₄ -H ₅)	(H ₆ -H ₈)	(O ₁)	(C ₁)	(C ₂)	(C ₃)

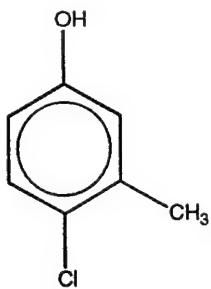
Probability:

I	II	III	IV	V	VI	VII	VIII
1/12	2/12	2/12	3/12	1/12	1/12	1/12	1/12

$$IC_2 = 5 \cdot 1/12 \cdot \log_2 12 + 2 \cdot 2/12 \cdot \log_2 12/2 + 3 \cdot 3/12 \cdot \log_2 12/3 = 2.855 \text{ bits}$$

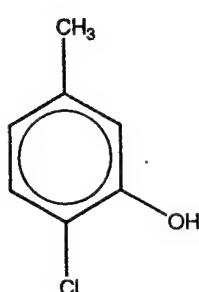
$$SIC_2 = IC_1 / \log_2 12 = 0.796 \text{ bits}$$

$$CIC_2 = \log_2 12 - IC_2 = 0.730 \text{ bits}$$



Probe: 3-methyl-4-chlorophenol

Structural:



(1) 0.00



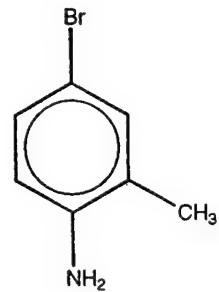
(2) 0.00



(3) 0.01

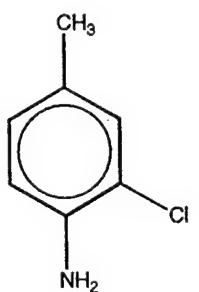


(4) 0.01

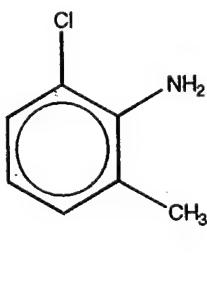


(5) 0.01

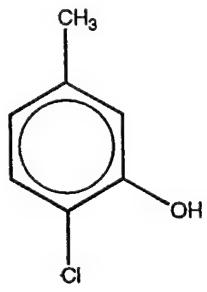
Chemical:



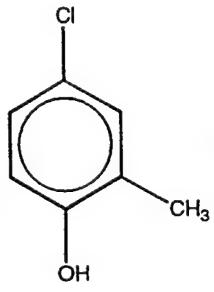
(1) 0.01



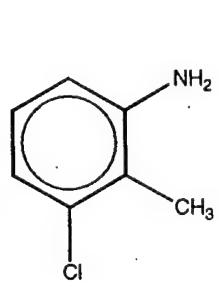
(2) 0.02



(3) 0.02

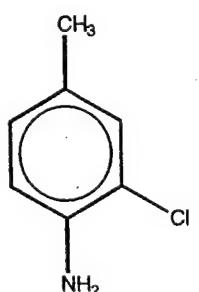


(4) 0.02

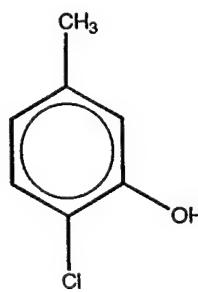


(5) 0.03

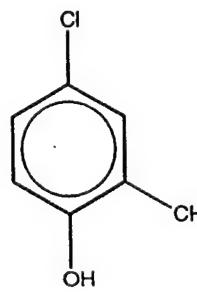
All:



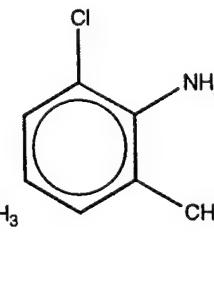
(1) 0.01



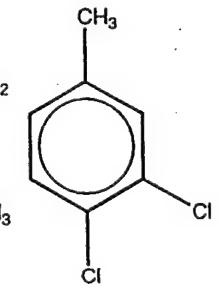
(2) 0.02



(3) 0.02



(4) 0.03



(5) 0.03

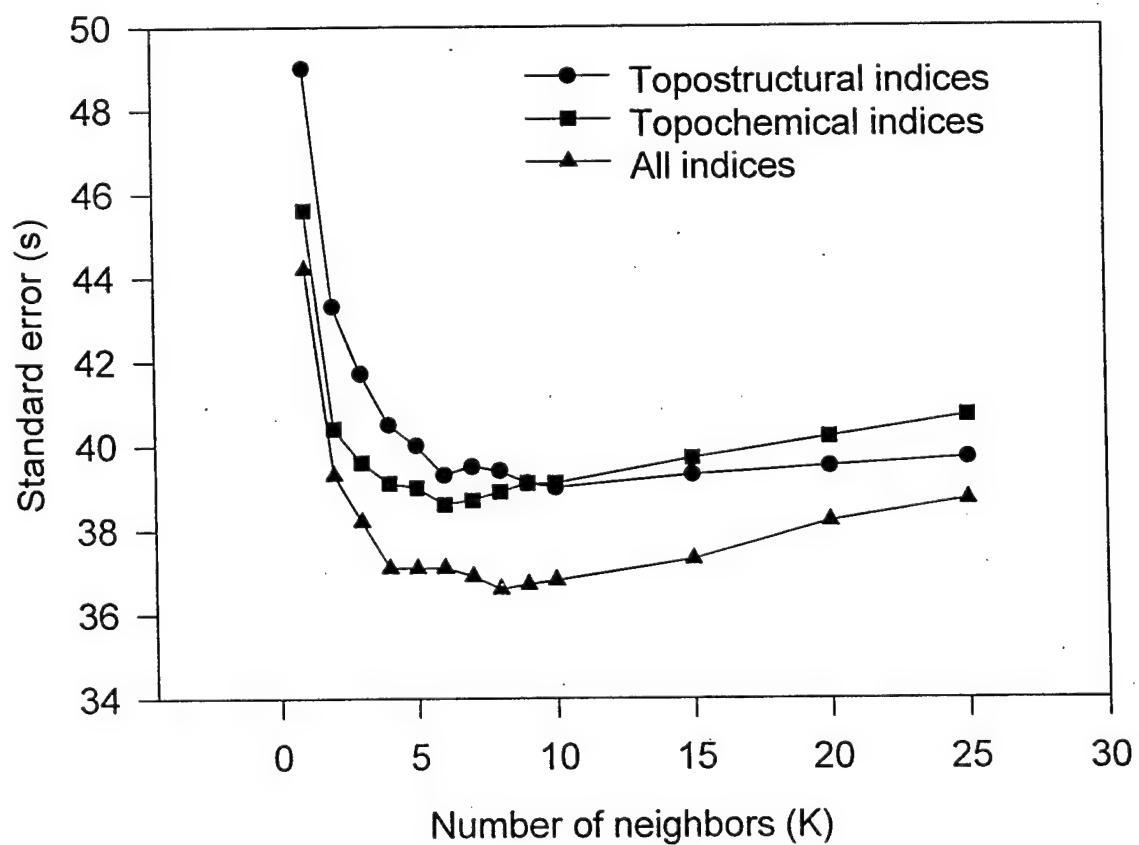
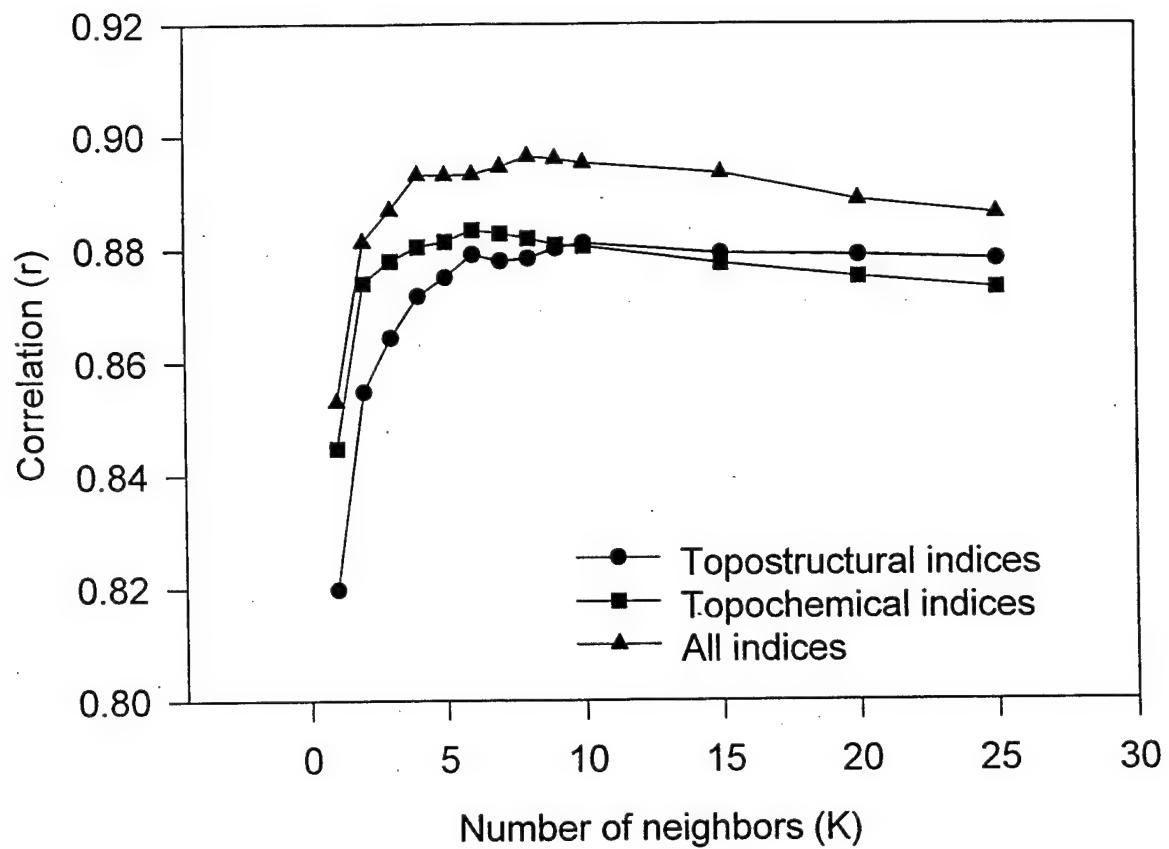


Table 1. Symbols, definitions and classifications of topological parameters.

Topostructural	
I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
\bar{I}_D^W	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
\bar{IC}	Information content of the distance matrix partitioned by frequency of occurrences of distance h
O	Order of neighborhood when IC reaches its maximum value for the hydrogen-filled graph
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
hX	Path connectivity index of order $h = 0-6$
hX	Cluster connectivity index of order $h = 3-6$
${}^hX_{PC}$	Path-cluster connectivity index of order $h = 4-6$
${}^hX_{Ch}$	Chain connectivity index of order $h = 3-6$
P_h	Number of paths of length $h = 0-10$
J	Balaban's J index based on distance
Topochemical	
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
IC_r	Mean information content or complexity of a graph based on the r^{th} ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r^{th} ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r^{th} ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph

${}^hX^b$	Bond path connectivity index of order $h = 0-6$
${}^hX_C^b$	Bond cluster connectivity index of order $h = 3-6$
${}^hX_{Ch}^b$	Bond chain connectivity index of order $h = 3-6$
${}^hX_{PC}^b$	Bond path-cluster connectivity index of order $h = 4-6$
${}^hX^v$	Valence path connectivity index of order $h = 0-6$
${}^hX_C^v$	Valence cluster connectivity index of order $h = 3-6$
${}^hX_{Ch}^v$	Valence chain connectivity index of order $h = 3-6$
${}^hX_{PC}^v$	Valence path-cluster connectivity index of order $h = 4-6$
J^B	Balaban's J index based on bond types
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii

Table 2. Summary of principal component analysis of 40 topostructural indices for 2926 chemicals.

PC	Eigenvalue	Proportion of explained variance	Cumulative explained variance	Top three correlated indices
1	28.2	46.2	46.2	$P_1, P_0, {}^1X$
2	11.0	18.0	64.3	${}^4X_{PC}, {}^5X_{PC}, {}^6X_{PC}$
3	5.9	9.6	73.9	${}^3X_C, {}^5X_C, {}^4X_{PC}$
4	4.1	6.7	80.6	$J, {}^6X_{Ch}, {}^4X_C$
5	2.8	4.6	85.2	${}^4X_{Ch}, {}^5X_{Ch}, {}^3X_{Ch}$
6	1.9	3.1	88.3	${}^3X_{Ch}, {}^4X_{Ch}, {}^5X_{Ch}$
7	1.5	2.4	90.8	${}^6X_C, P_{10}, P_9$

Table 3. Summary of principal component analysis of 61 topochemical indices for 2926 chemicals.

PC	Eigenvalue	Proportion of explained variance	Cumulative explained variance	Top three correlated indices
1	20.4	33.5	33.5	$^1X^b, ^2X^b, ^3X^b$
2	10.8	17.8	51.2	SIC_4, SIC_3, SIC_5
3	8.1	13.3	64.6	$^3X_C^b, ^4X_C^b, ^4X_{PC}^b$
4	6.1	9.9	74.5	$^5X_{Ch}^b, ^5X_{Ch}^v, ^4X_{Ch}^b$
5	3.0	5.0	79.5	$^3X_{Ch}^b, ^3X_{Ch}^v, ^4X_{Ch}^b$
6	2.4	3.9	83.4	IC_0, SIC_0, IC_1
7	1.7	2.8	86.2	$^6X_C^b, ^5X_C^b, ^6X_C^v$
8	1.4	2.2	88.4	$^4X_C^v, ^2X_C^v, ^6X_C^v$
9	1.2	2.0	90.4	$^5X_C^v, ^6X_C^v, ^4X_C^b$
10	1.1	1.8	92.1	$^4X_C^b, ^4X_C^v, ^6X_{PC}^v$

A Comparative Study of Topological and Geometrical Parameters in Estimating Normal Boiling Point and Octanol/Water Partition Coefficient

Subhash C. Basak, Brian D. Gute, and Gregory D. Grunwald

Natural Resources Research Institute, University of Minnesota,
Duluth, Duluth, Minnesota 55811

Journal of
**Chemical
Information and
Computer Sciences®**

Reprinted from
Volume 36, Number 6, Pages 1054-1060

A Comparative Study of Topological and Geometrical Parameters in Estimating Normal Boiling Point and Octanol/Water Partition Coefficient

Subhash C. Basak,* Brian D. Gute, and Gregory D. Grunwald

Natural Resources Research Institute, University of Minnesota, Duluth, Duluth, Minnesota 55811

Received February 22, 1996*

We have used topological, topochemical and geometrical parameters in predicting: (a) normal boiling point of a set of 1023 chemicals and (b) lipophilicity ($\log P$, octanol/water) of 219 chemicals. The results show that topological and topochemical variables can explain most of the variance in the data. The addition of geometrical parameters to the models provide marginal improvement in the model's predictive power. Among the three classes of descriptors, the topochemical indices were the most effective in predicting properties.

1. INTRODUCTION

A contemporary trend in theoretical chemistry, biomedical chemistry, drug design, and toxicology is the prediction of relevant properties of chemicals using structure-activity relationships (SARs).¹⁻⁹ A large number of SARs published in recent literature use parameters which can be calculated directly from molecular structure, as opposed to experimentally derived properties or parameters.^{1-6,10-30} The principal motivating factor behind this trend is our need to know many properties of a very large number of chemicals, both for practical drug design and hazard assessment of chemicals.^{13,31} All these properties cannot be determined experimentally due to limited resources. The modeling of the properties of chemicals using SARs based on calculated molecular descriptors has the following three major components:^{13,32}

1. Optimal representation of the chemical species by a chosen model object (structure representation).
2. Enumeration of relevant characteristics of the model object (parameterization).
3. Development of qualitative or quantitative models to predict properties using the selected structural characteristics (property prediction).

The first step in the overall process is representation (Figure 1). The term molecular structure represents a set of nonequivalent concepts. There is no reason to believe that when discussing different topics, e.g., chemical synthesis, reaction rates, spectroscopic transitions, reaction mechanisms, and ab initio calculations, that the term "molecular structure" represents the same fundamental reality.^{13,33} In fact, the various models of chemicals, e.g., classical valence bond representation, different graph theoretic representations, ball and spoke model of molecules, minimum energy conformation, and symbolic representation of molecules by Hamiltonian operators, are nothing but various representations of the same chemical entity. Once the model object is chosen, subsequent processes of parameterization and property estimation can be done in more than one way. Consequently, the field of theoretical SAR is comprised of a set of diverse modeling activities.

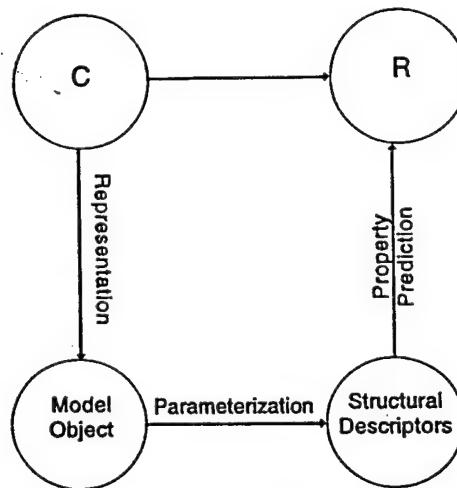


Figure 1. The processes of experimental determination vis-a-vis theoretical prediction of properties from SARs. C represents the set of chemicals and R the set of real numbers.

A convenient method of representing chemical species is by means of molecular graphs, where atoms are represented by vertices and bonds are depicted by edges.³⁴ Invariants derived from graphs can be used to characterize chemical structure. When a molecule is represented by a simple planar graph which does not distinguish among atoms or bond types, such invariants quantify molecular topology without being sensitive to such important chemical features like presence of heteroatoms or bond multiplicity. Such invariants may be termed "topological". On the other hand, when molecules are represented by graphs which are properly weighted to represent heterogeneity of atom types and bonding pattern, invariants derived from such graphs are chemically more realistic.³⁵ Such invariants have been found to be more useful as compared to the topological indices. We call such indices "topochemical" parameters, because they quantify both topology (connectivity) of atoms as well as the chemical characteristics of the specific molecular structure.

Another set of descriptors which have been used in many SARs are the geometrical or shape parameters, which encode information about the spatial characteristics of atoms in the molecule.³⁶⁻³⁸

In practical drug design and hazard assessment, where it is necessary to carry out very rapid estimation of a large

* All correspondence to be addressed to: Dr. Subhash C. Basak, Natural Resources Research Institute, University of Minnesota, Duluth, Duluth, MN 55811

Abstract published in *Advance ACS Abstracts*, October 1, 1996.

number of properties with no or very little empirical input, SARs based on topological, topochemical, and geometrical parameters can be of practical use. Therefore, in this paper, we have carried out a comparative study of topological, topochemical, and geometrical parameters in estimating (a) boiling point of a subset of the Toxic Substances Control Act (TSCA) Inventory comprising 1023 molecules and (b) lipophilicity of a set of 219 diverse compounds. The results are presented here with an analysis of the relative contributions of the three classes of indices in the development of SAR models.

2. MATERIALS AND METHODS

2.1. Normal Boiling Point Database. We used a subset of the Toxic Substances Control Act (TSCA) Inventory³¹ for which measured normal boiling point values were available and where HB_1 was equal to zero. HB_1 is a measure of the hydrogen bonding potential of a chemical. There were 1023 chemicals in the TSCA Inventory which satisfied these two criteria. Because of the large number of chemicals in this study, we are not listing the data for these chemicals in this paper. An electronic copy of the data may be obtained by contacting the authors.

2.2. Log *P* Database. Measured values of $\log P$ were obtained from CLOGP,³⁹ namely, the STARLIST group of chemicals. For this study, we used only chemicals where HB_1 was equal to zero. Also, the range of $\log P$ values for the purpose of estimation was restricted to -2 to 5.5. Actual measurements for $\log P$ beyond this range have been shown to be problematic.¹⁴ Table 1 provides a listing of the 219 chemicals that met these conditions.

2.3. Calculation of Topological and Geometric Parameters. Most of the topological indices used for property estimation were calculated by the computer program POLLY.⁴⁰ These indices include the molecular connectivity indices developed by Randić¹⁸ and Kier and Hall,³⁵ Wiener number,⁴¹ and frequency of path lengths of varying size. Information theoretic indices defined on the hydrogen-filled and hydrogen-suppressed molecular graph were calculated by POLLY using the methods of Basak *et al.*,^{42,43} Roy *et al.*,⁴⁴ Raychaudhury *et al.*,⁴⁵ and Bonchev and Trinajstić.⁴⁶ The *J* indices of Balaban⁴⁷⁻⁴⁹ were calculated using software developed by the authors. The hydrogen bonding parameter, HB_1 , was calculated using a program developed by Basak⁵⁰ and is based on the ideas of Ou *et al.*⁵¹

van der Waal's volume (V_w) was calculated using Sybyl 6.2.⁵² The 3-D Wiener numbers³⁷ were calculated using Sybyl with an SPL (Sybyl Programming Language) program developed by the authors. The calculation of the 3-D Wiener number consists of summing the entries in the upper triangular submatrix of the topographic Euclidean distance matrix for a molecule. The 3-D coordinates of each atom, needed for these computations, was determined using CONCORD 3.2.1.⁵³ For this paper, two variants of the 3-D Wiener number have been calculated, ^{3D}W and $^{3D}W_H$, where the hydrogen atoms have been excluded and included in the calculation, respectively.

In Table 2, the symbols for all topological and geometric parameters have been listed. A brief definition of each parameter is provided in Table 2 as well.

The parameters in Table 2 were then classified as being topological, topochemical, or geometric. Table 2 is orga-

nized to show where each parameter was classified. The topological parameters consist of those indices in which atom specific information and bonding type are ignored in calculation of the index. The topochemical indices account for atom and bond type information. The geometric parameters are based upon 3-D coordinate information of the molecule.

2.4. Statistical Analyses. Since the difference in magnitude for the topological and topochemical indices can vary greatly, they were transformed by the natural logarithm of the index plus one. One was added since many of the indices can be zero. The geometric parameters were transformed by the natural logarithm of the parameter.

Two regression procedures were used in the development of models. When the number of independent variables was high, typically greater than 25, a stepwise regression procedure to maximize improvement to R^2 was used. When the number of independent variables was small, all possible subsets regression was used. All regression models were developed using procedure REG of the SAS statistical package.⁵⁴

For both data sets, we randomly split the chemicals into approximately equal (50%/50%) training and test sets. For the BP data, there were 512 chemicals in the training set and 511 chemicals in the test set. For $\log P$, there were 114 chemicals in the training set and 105 chemicals in the test set. The training set and test set of chemicals are identified in Table 1 for the $\log P$ data. Models were developed using the training set of chemicals. These models were then used to predict the property values of the test chemicals. Final models were then developed using the combined training and test set of chemicals.

Initial models for the dependent property (BP or $\log P$) were developed using only the topological class of indices. Once the best topological model was determined, the topological indices used in the model were added to the set of topochemical indices. Then the best model from this combined set of indices was determined. Finally, the topological and/or topochemical indices used in the best model so far were added to the set of geometric parameters, and the best model using all of these parameters was determined.

3. RESULTS

3.1. TSCA Boiling Point Estimation. Stepwise regression analyses for BP of the training set of chemicals is summarized in Table 3. As is shown in Table 3, the topological model using 11 parameters resulted in an explained variance (R^2) of 80.8% and standard error (s) of 40.9 °C. Addition of the topochemical parameters with the 11 topological parameters increased the effectiveness of the model significantly. The resulting model used nine parameters, two topological parameters, and seven topochemical parameters. This model had an R^2 of 96.5% and s of 17.4 °C. All subsets regression of the nine topological and topochemical parameters retained thus far and the three geometric parameters resulted in a ten parameter model. This model included the nine topological and topochemical parameters and the geometric parameter $^{3D}W_H$. This model represented a slight improvement with R^2 of 96.7% and s of 16.8 °C.

Application of the three models to the test set of chemicals resulted in comparable R^2 and s and are listed in Table 3.

Table 1. Observed and Estimated Lipophilicity ($\log P$, Octanol/Water) for 219 Chemicals with HB_1 Equal to Zero

no.	chemical name	obs	est	est	est	no.	chemical name	obs	est	est	est
		$\log P$ (eq 4)	$\log P$ (eq 5)	$\log P$ (eq 6)	$\log P$ (eq 4)			$\log P$ (eq 4)	$\log P$ (eq 5)	$\log P$ (eq 6)	$\log P$ (eq 6)
1 ^a	1,4-dimethylnaphthalene	4.37	4.25	4.37	4.41	74 ^a	1,2,4-trichlorobenzene	4.02	3.65	3.84	3.79
2	cyclopropane	1.72	1.26	0.83	0.82	75 ^a	2,2',6-pcb	5.48	4.89	5.07	5.09
3	3,4-dimethylchlorobenzene	3.82	3.65	3.68	3.74	76	2-butyne	1.46	2.22	2.49	2.46
4	2,2-diphenyl-1,1,1-trichloroethane	4.87	4.90	4.93	4.99	77	azulene	3.20	3.59	3.52	3.45
5	2,6-dimethylnaphthalene	4.31	4.15	4.24	4.27	78 ^a	trifluoromethylthiobenzene	3.57	3.56	2.91	2.93
6	hexafluoroethane	2.00	2.63	2.59	2.33	79 ^a	2,5-pcb	5.16	4.62	4.90	4.89
7 ^a	1-iodoheptane	4.70	4.04	4.27	4.24	80 ^a	1,2,3-trichlorocyclohexene(34)	2.84	3.60	3.57	3.58
8 ^a	allyl bromide	1.79	2.22	2.04	2.06	81	biphenyl	4.09	4.18	4.33	4.32
9 ^a	1,5-dimethylnaphthalene	4.38	4.23	4.38	4.41	82 ^a	p-xylene	3.15	3.45	3.37	3.42
10	1,8-dimethylnaphthalene	4.26	4.31	4.41	4.43	83 ^a	ethylene	1.13	0.70	0.93	0.97
11 ^a	1,2,3-trichlorobenzene	4.05	3.60	3.64	3.63	84	thiophenol	2.52	3.08	3.01	3.04
12 ^a	2-ethylthiophene	2.87	3.20	2.69	2.73	85 ^a	bromotrifluoromethane	1.86	2.20	2.12	1.97
13	methylchloride	0.91	0.70	0.86	0.79	86	9-methylanthracene	5.07	5.07	4.90	4.92
14	γ -phenylpropylfluoride	2.95	3.73	3.26	3.29	87 ^a	trichloroethylene	2.42	2.63	2.44	2.44
15	iodobenzene	3.25	3.08	3.68	3.68	88 ^a	1,4-dimethyltetrachlorocyclohexane	4.40	4.18	4.03	4.11
16	1-methylpentachlorocyclohexane	4.04	4.18	4.20	4.24	89	propylene	1.77	1.38	1.59	1.71
17 ^a	ethane	1.81	0.70	1.50	1.47	90	cyclohexene	2.86	2.55	2.72	2.74
18	2,3'-pcb	5.02	4.71	4.99	4.97	91 ^a	methylthiobenzene	2.74	3.28	3.02	2.97
19	cyclopentane	3.00	2.19	2.35	2.37	92	methylfluoride	.51	0.70	0.57	0.53
20	ethylchloride	1.43	1.38	1.48	1.52	93	γ -phenylpropyl iodide	3.90	3.73	4.06	4.06
21	2-phenylthiophene	3.74	3.88	4.01	3.96	94	2,3,4'-pcb	5.42	4.89	5.10	5.10
22	trichlorofluoromethane	2.53	2.20	2.34	2.29	95 ^a	fluoropentachlorocyclohexane	3.19	4.18	3.87	3.89
23 ^a	fluoroform	0.64	1.85	0.57	0.42	96	1,2,3,5-tetrachlorobenzene	4.92	3.91	4.08	4.04
24	dimethyl disulfide	1.77	2.22	1.57	1.39	97	2,2'-pcb	4.90	4.65	4.80	4.82
25 ^a	propane	2.36	1.38	1.97	2.01	98	1-butene	2.40	2.22	1.96	2.05
26	hexamethylbenzene	5.11	4.18	4.94	4.97	99 ^a	1,3-dimethylnaphthalene	4.42	4.29	4.43	4.46
27	butanethiol	2.28	2.80	2.81	2.87	100 ^a	1,7-dimethylnaphthalene	4.44	4.23	4.43	4.45
28 ^a	diethylsulfide	1.95	2.80	2.68	2.67	101 ^a	1-methylnaphthalene	3.87	3.95	4.08	4.07
29	cyclohexane	3.44	2.55	2.83	2.87	102	2,6-pcb	4.93	4.70	4.83	4.85
30 ^a	diphenyl disulfide	4.41	4.62	4.57	4.53	103 ^a	α -bromotoluene	2.92	3.28	3.47	3.42
31	<i>m</i> -fluorobenzylchloride	2.77	3.55	2.95	2.99	104	2,2',3'-trichlorobiphenyl	5.31	4.89	5.22	5.20
32	1-chloropropane	2.04	2.22	1.92	1.97	105	hexafluorobenzene	2.22	4.18	3.20	2.97
33	2,4-dichlorobenzylchloride	3.82	4.01	3.67	3.69	106 ^a	3-bromothiophene	2.62	2.49	2.73	2.72
34	<i>m</i> -chlorotoluene	3.28	3.58	3.30	3.34	107 ^a	1,2,3,5-tetramethylbenzene	4.17	3.91	4.27	4.32
35 ^a	butane	2.89	2.22	2.39	2.43	108	halothane	2.30	3.01	2.16	2.19
36	1,2,3-trimethylbenzene	3.66	3.60	3.89	3.93	109	2,4,6-pcb	5.47	4.97	5.02	5.04
37	1,1-difluoroethylene	1.24	1.85	0.72	0.79	110	1,1-dichloroethylene	2.13	1.85	1.89	1.97
38 ^a	1-chlorobutane	2.64	2.80	2.67	2.70	111	α -dibromobenzene	3.64	3.29	3.94	3.88
39	2,3-dibromo thiophene	3.53	2.98	3.22	3.23	112	1,2,4,5-tetramethylbenzene	4.00	3.82	4.24	4.29
40 ^a	pentafluor ethylbenzene	3.36	3.24	3.15	3.18	113	1-hexene	3.39	3.25	3.17	3.21
41 ^a	1,2,4,5-tetrabromobenzene	5.13	3.82	5.06	4.94	114 ^a	neopentane	3.11	2.20	3.12	3.28
42	<i>o</i> -dichlorobenzene	3.38	3.29	3.19	3.19	115	chloroform	1.97	1.85	2.11	2.06
43 ^a	1,2,3,4-tetrachlorobenzene	4.64	3.81	4.01	3.97	116 ^a	1-fluorobutane	2.58	2.80	2.15	2.20
44 ^a	tribromoethene	3.20	2.63	3.37	3.26	117 ^a	pyrene	4.88	5.46	4.90	4.88
45	pentane	3.39	2.80	3.01	3.03	118	1,1-dichloro-2,2-diphenylethane	4.51	4.95	4.88	4.94
46 ^a	isobutane	2.76	1.85	2.61	2.71	119 ^a	isobutylene	2.34	1.85	2.47	2.61
47 ^a	mirex	5.28	5.10	5.26	5.18	120	diphenylmethane	4.14	4.40	4.51	4.54
48 ^a	1,3-dichlorobenzene	3.60	3.58	3.24	3.23	121	isopropylbenzene	3.66	3.35	3.62	3.67
49 ^a	1,2-dimethylnaphthalene	4.31	4.25	4.41	4.43	122 ^a	naphthalene	3.30	3.43	3.38	3.34
50 ^a	2-ethyl naphthalene	4.38	4.32	4.23	4.24	123 ^a	1-heptene	3.99	3.86	3.48	3.50
51 ^a	cycloheptatriene	2.63	3.59	2.74	2.74	124	2,2-dimethylbutane	3.82	2.88	3.45	3.55
52 ^a	3-chlorobiphenyl	4.58	4.42	4.65	4.64	125	1-fluoropentane	2.33	3.25	2.79	2.82
53 ^a	3-ethylthiophene	2.82	3.20	2.72	2.75	126 ^a	<i>o</i> -xylene	3.12	3.29	3.44	3.49
54	1,3,5-tribromobenzene	4.51	4.00	4.54	4.48	127 ^a	ethylbenzene	3.15	3.28	3.25	3.26
55 ^a	β -phenylethylchloride	2.95	3.50	3.31	3.33	128 ^a	trichloromethylthiobenzene	3.78	3.56	3.59	3.61
56	acenaphthene	3.92	4.49	3.94	3.95	129 ^a	thiophene	1.81	2.19	1.64	1.62
57	<i>m</i> -dibromobenzene	3.75	3.58	4.06	3.98	130	bromo-chloromethane	1.41	1.38	1.49	1.47
58	dichlorodifluoromethane	2.16	2.20	1.88	1.83	131 ^a	1,2-dichlorotetrafluoroethane	2.82	2.63	2.71	2.65
59	toluene	2.73	3.08	3.04	3.05	132 ^a	2-chlorobiphenyl	4.38	4.43	4.65	4.65
60 ^a	anthracene	4.45	4.85	4.62	4.59	133	2,4'-dichlorobiphenyl	5.10	4.68	4.88	4.87
61 ^a	hexachlorocyclopentadiene	5.04	4.00	4.99	4.86	134 ^a	1,3,5-trichlorobenzene	4.15	4.00	3.48	3.50
62	3-phenyl-1-chloropropane	3.55	3.73	3.56	3.58	135	1-octene	4.57	4.04	3.77	3.78
63 ^a	biphenyl	4.79	4.62	4.69	4.71	136	methylbromide	1.19	0.70	1.23	1.07
64 ^a	1-chloroheptane	4.15	4.04	3.72	3.71	137 ^a	phenylethylsulfide	3.20	3.50	3.37	3.36
65 ^a	2,4-dichlorotoluene	4.24	3.65	3.60	3.64	138	1-ethyl-2-methylbenzene	3.53	3.54	3.81	3.84
66 ^a	1,1-dichloroethane	1.79	1.85	1.93	2.02	139 ^a	propylbenzene	3.72	3.50	3.56	3.58
67 ^a	(β)-benzothiophene	3.12	3.15	3.24	3.17	140 ^a	indane	3.18	3.15	3.06	3.04
68 ^a	2-bromo thiophene	2.75	2.49	2.62	2.61	141	2-chloropropane	1.90	1.85	2.22	2.33
69	chlorodifluoromethane	1.08	1.85	0.75	0.75	142 ^a	phenylazide	2.59	3.50	2.83	2.88
70 ^a	pentachlorobenzene	5.17	4.02	4.62	4.52	143	2,4-dibromotetrafluorocyclohexane	3.98	4.18	4.25	4.29
71	9,10-dihydroanthracene	4.25	4.85	4.31	4.34	144 ^a	tertachloroethylene	3.40	3.01	3.69	3.52
72	1,3-(bis-chloromethyl)benzene	2.72	3.85	3.49	3.51	145	1-nonene	5.15	4.27	3.97	3.98
73	chlorobenzene	2.84	3.08	2.90	2.90	146	2,3-dimethylbutane	3.85	3.01	3.41	3.50

Table 1 (Continued)

no.	chemical name	obs log P (eq 4)	est log P (eq 5)	est log P (eq 6)	no.	chemical name	obs log P (eq 4)	est log P (eq 5)	est log P (eq 6)		
			log P (eq 4)	log P (eq 5)				log P (eq 4)	log P (eq 5)		
147 ^a	dichlorofluoromethane	1.55	1.85	1.25	1.30	184 ^a	2,3,6-trimethylnaphthalene	4.73	4.46	4.61	4.64
148 ^a	1,1,2,2-tetrachloroethane	2.39	3.01	2.91	2.90	185 ^a	difluoromethane	.20	1.38	0.20	0.11
149 ^a	1,2,4-trimethylbenzene	3.78	3.65	3.95	3.98	186	1,2,4-trifluorobenzene	2.52	3.65	2.63	2.55
150 ^a	fluorobenzene	2.27	3.08	2.39	2.40	187	bromobenzene	2.99	3.08	3.38	3.33
151	butylbenzene	4.26	3.73	3.81	3.83	188	hexachloro-1,3-butadiene	4.78	4.26	5.00	4.86
152 ^a	ethylbromide	1.61	1.38	1.98	1.95	189	vinylbromide	1.57	1.38	1.76	1.78
153 ^a	tetrafluoromethane	1.18	2.20	1.61	1.29	190 ^a	o-chlorotoluene	3.42	3.29	3.33	3.36
154 ^a	p-cymene	4.10	3.93	3.88	3.92	191 ^a	α-chlorotoluene	2.30	3.28	3.09	3.10
155 ^a	p-chlorotoluene	3.33	3.45	3.17	3.22	192	1,4-cyclohexadiene	2.30	2.55	2.42	2.46
156 ^a	1-bromopropane	2.10	2.22	2.35	2.34	193	1-bromoheptane	4.36	4.04	4.06	4.00
157 ^a	bromocyclohexane	3.20	3.08	3.47	3.46	194	styrene	2.95	3.28	3.15	3.17
158 ^a	2-methylthiophene	2.33	2.49	2.39	2.41	195	chlorotrifluoromethane	1.65	2.20	1.56	1.45
159	diphenylsulfide	4.45	4.40	4.48	4.47	196 ^a	(dimethyl)phenylphosphine	2.57	3.35	2.99	2.92
160 ^a	1,2,4,5-tetrachlorobenzene	4.82	3.82	3.93	3.91	197	cycloocta-1,5-diene	3.16	3.23	2.88	2.93
161	1,1,1-trichloroethane	2.49	2.20	2.43	2.52	198	tetrachlorocyclohexane	2.82	3.82	3.50	3.55
162 ^a	p-dichlorobenzene	3.52	3.45	3.08	3.10	199	1-bromooctane	4.89	4.27	4.23	4.17
163	1-bromobutane	2.75	2.80	3.15	3.14	200	2-methylnaphthalene	3.86	3.90	4.03	4.01
164 ^a	p-chlorobiphenyl	4.61	4.50	4.57	4.56	201	3-methylthiophene	2.34	2.49	2.44	2.46
165 ^a	cyclopropylbenzene	3.27	2.98	3.01	3.02	202 ^a	methylenechloride	1.25	1.38	1.38	1.35
166 ^a	2,6-dichlorotoluene	4.29	3.60	3.48	3.54	203 ^a	hexachlorobenzene	5.31	4.18	5.09	4.93
167 ^a	allene	1.45	1.38	1.42	1.48	204	indene	2.92	3.15	3.01	3.01
168 ^a	b-phenylethylbromide	3.09	3.50	3.66	3.64	205	tert-butylbenzene	4.11	3.26	3.92	3.99
169 ^a	1,3-butadiene	1.99	2.22	1.88	1.97	206	1,2-dichloroethane	1.48	2.22	1.92	1.91
170	2-chlorothiophene	2.54	2.49	2.14	2.16	207 ^a	1,3,5-trimethylbenzene	3.42	4.00	3.81	3.87
171	1-bromopentane	3.37	3.25	3.62	3.58	208 ^a	phenanthrene	4.46	4.88	4.69	4.68
172 ^a	γ-phenylpropylbromide	3.72	3.73	3.87	3.84	209 ^a	benzene	2.13	2.55	2.40	2.39
173	1,3-cyclohexadiene	2.47	2.55	2.47	2.50	210	3,3,3-trifluoropropylbenzene	3.31	3.80	3.19	3.24
174 ^a	pentamethylbenzene	4.56	4.02	4.65	4.69	211 ^a	α-(2,2,2-trichloroethyl)styrene	4.56	3.93	4.04	4.13
175 ^a	p-dibromobenzene	3.79	3.45	3.81	3.76	212 ^a	2,3-dimethylnaphthalene	4.40	4.20	4.25	4.28
176	1,4-pentadiene	2.48	2.80	2.32	2.42	213 ^a	1,3-dichloropropane	2.00	2.80	2.47	2.47
177 ^a	methyliodide	1.51	0.70	1.48	1.42	214	1,2,3,4-tetramethylbenzene	4.11	3.81	4.26	4.31
178 ^a	1,1-difluoroethane	.75	1.85	1.04	1.11	215 ^a	stilbene-t	4.81	4.62	4.79	4.78
179 ^a	1-bromohexane	3.80	3.86	3.80	3.75	216	fluorene	4.18	4.65	4.22	4.21
180 ^a	m-xylene	3.20	3.58	3.44	3.49	217 ^a	2-fluoro-3-bromotetrachlorocyclohexane	3.28	4.18	4.06	4.09
181	dibenzothiophene	4.38	4.65	4.44	4.40	218 ^a	allylbenzene	3.23	3.50	3.37	3.41
182	ethyliodide	2.00	1.38	2.28	2.34	219 ^a	carbontetrachloride	2.83	2.20	3.27	3.10
183	trifluoromethylbenzene	3.01	3.26	2.77	2.80						

^a Training chemicals.

The largest difference in variance explained was for the topological parameter model. For this model, R^2 decreased from 80.8% to 79.5% or 1.3% less variance explained. However, the standard error for the test chemicals was 0.1 °C lower. For the other two models, the R^2 of the test chemicals was within 0.6% of that seen for the training chemicals. Standard errors for the test chemicals were within 1 °C of the standard error for the training set of chemicals.

Regression analysis of the set of training and test chemicals combined showed similar results as analysis of the training set of chemicals. Using only the topological class of indices, stepwise regression resulted in an eight parameter model to estimate boiling point:

$$\text{BP} = -21.9 + 30.6(W) - 21.5(O) + 69.9(^6\chi) + 35.8(^6\chi) - 106.5(^6\chi_C) - 96.1(^5\chi_{\text{CH}}) - 17.7(^5\chi_{\text{PC}}) + 19.5(P_{10}) \quad (1)$$

$$n = 1023, R^2 = 81.2\%, s = 39.7 \text{ °C}, F = 547$$

These eight parameters were added to the set of topochemical parameters. Again, stepwise regression was used to develop a model using the eight topological and all topochemical indices. The best model to estimate boiling

point consisted of eight parameters again:

$$\text{BP} = -332.9 + 134.6(^6\chi) + 10.9(P_{10}) + 110.0(\text{IC}_0) - 133.8(^6\chi) - 80.2(^3\chi_C) + 176.5(^0\chi) + 44.8(^2\chi) + 16.8(^5\chi_{\text{PC}}) \quad (2)$$

$$n = 1023, R^2 = 96.1\%, s = 18.0 \text{ °C}, F = 3151$$

Only two of the topological indices used in eq 1 were retained by the regression procedure in eq 2: $^6\chi$ and P_{10} . The improvement in R^2 was very significant, going from 81.2% for eq 1 to 96.1% for eq 2. Also, the model error decreased by over half, dropping from 39.7 °C to 18.0 °C.

Using all subsets regression on the eight parameters of eq 2 and the three geometric parameters resulted in a ten parameter model as follows:

$$\text{BP} = -285.7 + 125.3(^6\chi) + 10.6(P_{10}) + 74.5(\text{IC}_0) - 125.0(^6\chi) - 86.3(^3\chi_C) + 175.3(^0\chi) + 49.1(^2\chi) + 18.7(^5\chi_{\text{PC}}) - 9.1(^3\text{D}W_H) + 8.1(^3\text{D}W) \quad (3)$$

$$n = 1023, R^2 = 96.3\%, s = 17.6 \text{ °C}, F = 2650$$

Equation 3 contains all of the parameters from eq 2 plus the two variants of the 3-D Wiener number. The addition

Table 2. Symbols, Definitions, and Classifications of Topological and Geometrical Parameters

Topological	
I_D^W	information index for the magnitudes of distances between all possible pairs of vertices of a graph
I_D^W	mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^P	degree complexity
H^V	graph vertex complexity
H^D	graph distance complexity
IC	information content of the distance matrix partitioned by frequency of occurrences of distance h
O	order of neighborhood when IC reaches its maximum value for the hydrogen-filled graph
M_1	a Zagreb group parameter = sum of square of degree over all vertices
M_2	a Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
$h\chi$	path connectivity index of order $h = 0-6$
$h\chi$	cluster connectivity index of order $h = 3-5$
$h\chi_{PC}$	path-cluster connectivity index of order $h = 4-6$
$h\chi_{Ch}$	chain connectivity index of order $h = 5-6$
P_h	number of paths of length $h = 0-10$
J	Balaban's J index based on distance
Topochemical	
I_{ORB}	information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
IC_r	mean information content or complexity of a graph based on the r th ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	structural information content for r th ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	complementary information content for r th ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
$h\chi^b$	bond path connectivity index of order $h = 0-6$
$h\chi^b_C$	bond cluster connectivity index of order $h = 3-5$
$h\chi^b_{Ch}$	bond chain connectivity index of order $h = 5-6$
$h\chi^b_{PC}$	bond path-cluster connectivity index of order $h = 4-6$
$h\chi^v$	valence path connectivity index of order $h = 0-6$
$h\chi^v_C$	valence cluster connectivity index of order $h = 3-5$
$h\chi^v_{Ch}$	valence chain connectivity index of order $h = 5-6$
$h\chi^v_{PC}$	valence path-cluster connectivity index of order $h = 4-6$
J^B	Balaban's J index based on bond types
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii
Geometric	
V_W	van der Waal's volume
$3DW$	3-D Wiener number for the hydrogen-suppressed geometric distance matrix
$3DW_H$	3-D Wiener number for the hydrogen-filled geometric distance matrix

Table 3. Summary of Regression Results for the Training Set of Chemicals and Predictions of Test Set of Chemicals for Dependent Variable BP (°C) for Three Parameter Classes

parameter class	training set ($N = 512$)			test set ($N = 511$)	
	variables included	F	R^2	s	R^2
topological	$\overline{IC}, O, M_2, {}^3\chi, {}^3\chi_c, {}^3\chi_{C_6}, {}^5\chi_{C_6}, {}^5\chi_{C_6}, P_{10}, J$	191	80.8	40.9	79.5
topological + topochemical	$\overline{\chi}, P_{10}, IC_0, {}^3\chi, {}^3\chi_c, {}^3\chi_{C_6}, {}^3\chi_{C_6}, {}^3\chi_{PC}$	1547	96.5	17.4	96.0
topological + topochemical + geometric	$\overline{\chi}, P_{10}, IC_0, {}^3\chi, {}^3\chi_c, {}^3\chi_{C_6}, {}^3\chi_{C_6}, {}^3\chi_{PC}, {}^{3D}W_H$	1486	96.7	16.8	96.1

Table 4. Summary of Regression Results for the Training Set of Chemicals and Predictions of Test Set of Chemicals for Dependent Variable Log *P* for Three Parameter Classes

of the two 3D-Wiener numbers resulted in only a very slight increase in the predictive power of the model. The standard error (s) decreased by only $0.4\text{ }^{\circ}\text{C}$ with the addition of the geometric parameters and R^2 increased from 96.1% to 96.3%, an increase of only 0.2% of the variance explained by eq 2 over eq 3. A scatterplot of observed boiling point *vs* estimated boiling point using eq 3 is shown in Figure 2.

3.2. Log *P* Estimation. Stepwise regression analyses for log *P* of the training set of chemicals is summarized in Table 4. The topological parameter model included nine variables. These nine variables explained 77.9% of the variance with

a standard error of 0.57. Regression analysis of these nine topological parameters and the topochemical parameters resulted in a better model with only seven parameters. This model included two topological parameters and five topochemical. The R^2 increased from 77.9% to 89.0% and s decreased from 0.57 to 0.40. Adding the geometric parameters provided a very minor increase. For this model, ^{3D}W replaced $^3\chi^b$, the R^2 remained the same, and s decreased from 0.40 to 0.39.

Application of these models to the test set of chemicals resulted in slightly decreased variance explained and slightly

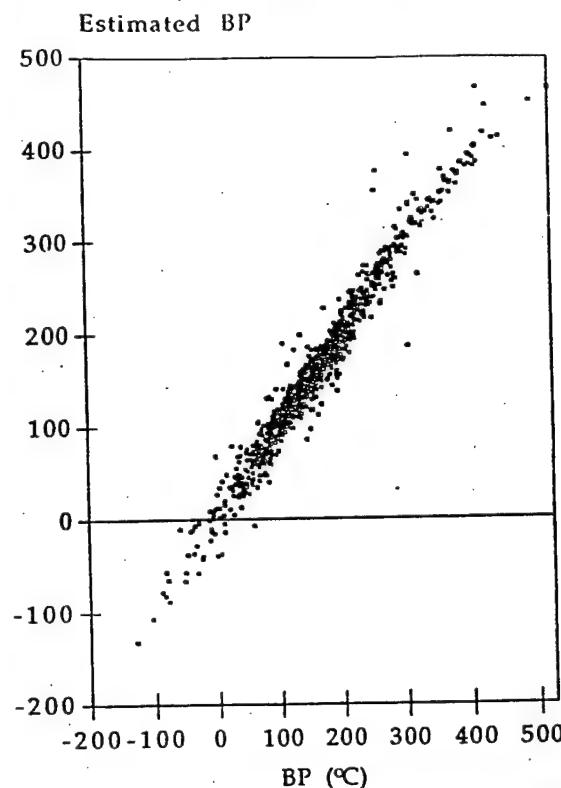


Figure 2. Scatterplot of observed boiling point *vs* estimated boiling point using eq 3 for 1023 diverse chemicals.

increased standard error. All R^2 for the test set differed by no more than 4.1% of the R^2 seen for the training set. The standard error of the test set of chemicals was within 0.06 of the standard error of the training set. These results can be seen in Table 4.

As with the BP data set, regression analyses of the combined training and test sets was similar to the analyses of the training sets. Starting with topological parameters only, the following seven parameter model was developed to estimate $\log P$:

$$\log P = -1.42 + 1.08(W) - 1.58(^2\chi) + 1.51(^6\chi) - 0.92(^6\chi_C) - 0.32(P_7) + 0.20(P_{10}) + 1.97(J) \quad (4)$$

$$n = 219, R^2 = 78.9\%, s = 0.54, F = 112$$

The seven parameters of eq 4 were added to the set of topochemical indices, and a new model was developed using stepwise regression. This new model consisted of ten parameters:

$$\log P = -2.13 - 0.20(^2\chi) + 0.18(P_{10}) - 1.86(IC_0) + 1.33(CIC_2) - 0.92(CIC_3) - 1.36(^6\chi) + 5.76(^0\chi^v) - 2.98(^1\chi^v) + 0.54(^4\chi^v) - 0.39(^3\chi^v_C) \quad (5)$$

$$n = 219, R^2 = 90.8\%, s = 0.36, F = 206$$

As with the boiling point models, only two of the topological parameters were retained in eq 5, $^2\chi$ and P_{10} . Also, just like the boiling point models, the addition of the topochemical parameters resulted in a significant increase in the quality of $\log P$ estimation.

All subsets regression using the ten parameters of eq 5 and the geometric parameters resulted in the following 11

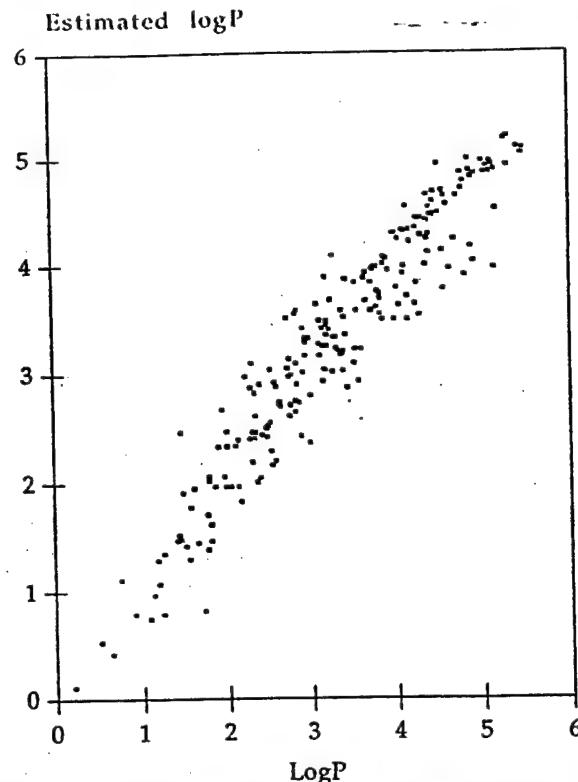


Figure 3. Scatterplot of observed $\log P$ *vs* estimated $\log P$ using eq 6 for 219 diverse chemicals.

parameter model:

$$\log P = -5.60 + 0.19(P_{10}) - 1.46(IC_0) + 1.09(CIC_2) - 0.77(CIC_3) - 1.36(^6\chi) + 5.34(^0\chi^v) - 3.41(^1\chi^v) + 0.55(^4\chi^v) - 0.41(^3\chi^v_C) + 1.10(V_W) - 0.17(^3D_W) \quad (6)$$

$$n = 219, R^2 = 91.2\%, s = 0.35, F = 194$$

Equation 6 differs from eq 5 with the removal of $^2\chi$ and the addition of V_W and 3D_W . The addition of the geometric parameters resulted in only slight improvement in the ability to estimate $\log P$.

Estimated $\log P$ values using eqs 4–6 may be found in Table 1. Figure 3 shows a scatterplot of observed $\log P$ *vs* estimated $\log P$ using eq 6.

4. DISCUSSION

The objective of this paper was to carry out a comparative study of the effectiveness of topological, topochemical, and geometrical parameters in SAR. To this end, we used these three classes of parameters in predicting normal boiling point of a diverse set of 1023 chemicals and $\log P$ of a set of 219 chemicals. To further assess the utility of these models for predictive purposes, the data sets were split into training and test sets by randomly assigning chemicals to one or the other. Models developed using the training sets of chemicals were used to predict the relevant property of the test chemicals.

As can be seen in Tables 3 and Table 4, the models developed using the training sets of chemicals could predict BP and $\log P$ of the test chemicals as accurately as they could estimate these properties for the training chemicals. Therefore, it seemed reasonable to combine the training and test sets to develop the regression models.

Both for boiling point and $\log P$, topological variables gave a reasonable predictive model. The addition of topochemical parameters to the set of independent variables resulted in substantial improvement in model performance. Further addition of geometrical variables gave slight improvement in explained variance in these data.

Our modeling approach in this paper was a hierarchical one, beginning with parameters derived from the simplest (topological) representation of molecules. Such indices are derived from simple graphs which are unweighted and, consequently, do not represent the reality of chemicals very well. The next tier of variables, topochemical indices, quantify information both about topology as well as atom types and bonding pattern. Finally, geometrical or 3-D parameters were used for modeling. The results show that the addition of chemical information makes a substantial contribution to the predictive power of the models for both boiling point and $\log P$. It would be interesting to see whether this trend is valid for other properties.

ACKNOWLEDGMENT

This is contribution number 183 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported in part by grant F49620-94-1-0401 from the United States Air Force, a grant from Exxon Corporation, and the Structure-Activity Relationship Consortium (SARCON) of the Natural Resources Research Institute of the University of Minnesota.

REFERENCES AND NOTES

- Wilkins, C. L.; Randić, M. *Theoret. Chim. Acta (Berl.)* 1980, 58, 45.
- Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* 1985, 25, 64.
- Johnson, M. A.; Basak, S. C.; Maggiore, G. *Mathl. Comput. Modeling* 1988, 11, 630.
- Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. *Discrete Appl. Math.* 1988, 19, 17.
- Basak, S. C.; Bertelsen, S.; Grunwald, G. D. *J. Chem. Inf. Comput. Sci.* 1994, 34, 270.
- Basak, S. C.; Grunwald, G. D. *Math. Modeling Sci. Comput.* 1994, in press.
- Willet, P.; Winterman, V. *Quant. Struct.-Act. Relat.* 1986, 5, 18.
- Fisanick, W.; Cross, K. P.; Rusinko, III, A. *J. Chem. Inf. Comput. Sci.* 1992, 32, 664.
- Lajiness, M. S. In *Computational Chemical Graph Theory*, Rouvray, D. H., Ed.; Nova Publishers: New York, 1990; p 299.
- Niemi, G. J.; Basak, S. C.; Veith, G. D. In *Proceedings of the First Conference of the International Society of Environmental Protection*, Vol. 2, No. 1; Zirm, K., Mayer, J., Eds.; WB. Druck Bmg. and Co.: Reiden, Austria, 1989; p 57.
- Basak, S. C.; Niemi, G. J.; Veith, G. D. *J. Math. Chem.* 1990, 4, 185.
- Basak, S. C.; Niemi, G. J.; Veith, G. D. *Mathematical and Computer Modeling* 1990, 14, 511.
- Basak, S. C.; Niemi, G. J.; Veith, G. D. *J. Math. Chem.*, 1991, 7, 243.
- Niemi, G. J.; Basak, S. C.; Veith, G. D.; Grunwald, G. D. *Environ. Toxicol. Chem.* 1992, 11, 893.
- Basak, S. C.; Grunwald, G. D. *Math. Modeling Sci. Comput.* 1993, 2, 735.
- Randić, M.; Basak, S. C. *Mathematical and Computer Modeling* (in preparation).
- Randić, M. *Int. J. Quantum Chem.* 1984, 11, 137.
- Randić, M. *J. Am. Chem. Soc.* 1975, 97, 6609.
- Randić, M. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiore, G. M., Eds.; Wiley: New York, 1990; p 77.
- Balasubramanian, K. *SAR QSAR Environ. Res.* 1994, 2, 59.
- Rouvray, D. H. *J. Mol. Struct. (Theochem)* 1995, 336, 101.
- Balaban, A. T.; Basak, S. C.; Colburn, T.; Grunwald, G. D. *J. Chem. Inf. Comput. Sci.* 1994, 34, 1118.
- Basak, S. C.; Grunwald, G. D. *SAR QSAR Environ. Res.* 1994, 2, 289.
- Basak, S. C.; Grunwald, G. D. *New J. Chem.* 1995, 19, 231.
- Basak, S. C.; Grunwald, G. D. In *Proceeding of the XVI International Cancer Congress*; Rao, R. S., Deo, M. G., Sanghvi, L. D., Eds.; Monduzzi: Bologna, Italy, 1995; p 413.
- Basak, S. C.; Grunwald, G. D. *J. Chem. Inf. Comput. Sci.* 1995, 35, 366.
- Basak, S. C.; Grunwald, G. D. *SAR QSAR Environ. Res.* 1995, 3, 265.
- Basak, S. C.; Grunwald, G. D. *Chemosphere* 1995, 31, 2529.
- Basak, S. C.; Bertelsen, S.; Grunwald, G. D. *Toxicology Lett.* 1995, 79, 239.
- Basak, S. C.; Gute, B. D.; Drewes, L. R. *Pharm. Res.* 1996, 13, 775.
- Auer, C. M.; Nabholz, J. V.; Baetcke, K. P. *Environ. Health Perspect.* 1990, 87, 183.
- Bunge, M. *Method, Model and Matter*; D. Reidel Publishing Company: Dordrecht-Holland/Boston, 1973.
- Weininger, S. *J. J. Chem. Educ.* 1984, 61, 939.
- Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1983.
- Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Research Studies Press: Chichester, England, 1986.
- Moriguchi, I.; Kanada, Y. *Chem. Pharm. Bull.* 1977, 25, 926.
- Bogdanov, B.; Nikolić, S.; Trinajstić, N. *J. Math. Chem.* 1989, 3, 299.
- Kamlet, M. J.; Doherty, R. M.; Abraham, M. H.; Marcus, Y.; Taft, R. W. *J. Phys. Chem.* 1988, 92, 5244.
- Leo, A.; Weininger, D. *CLOGP Version 3.2 User Reference Manual*; Medicinal Chemistry Project, Pomona College, Claremont, CA, 1984.
- Basak, S. C.; Harriss, D. K.; Magnuson, V. R. *POLLY Version 2.3*; Copyright of the University of Minnesota, 1988.
- Wiener, H. *J. Am. Chem. Soc.* 1947, 69, 17.
- Basak, S. C.; Roy, A. B.; Ghosh, J. J. In *Proceedings of the IIInd International Conference on Mathematical Modeling*; Avula, X. J. R., Bellman, R., Luke, Y. L., Rigler, A. K., Eds.; University of Missouri-Rolla: Rolla, MO, 1979; p 851.
- Basak, S. C.; Magnuson, V. R. *Arzneim.-Forsch./Drug Res.* 1983, 33, 501.
- Roy, A. B.; Basak, S. C.; Harriss, D. K.; Magnuson, V. R. In *Mathematical Modeling in Science and Technology*; Avula, X. J. R., Kalman, R. E., Lipaia, A. I., Rodin, E. Y., Eds.; Pergamon: New York, 1984; p 745.
- Raychaudhury, C.; Ray, S. K.; Ghosh, J. J.; Roy, A. B.; Basak, S. C. *J. Comput. Chem.* 1984, 5, 581.
- Bonchev, D.; Trinajstić, N. *J. Chem. Phys.* 1977, 67, 4517.
- Balaban, A. T. *Chem. Phys. Lett.* 1982, 89, 399.
- Balaban, A. T. *Pure Appl. Chem.* 1982, 55, 199.
- Balaban, A. T. *MATCH* 1986, 21, 115.
- Basak, S. C. *H-Bond*; Copyright of the University of Minnesota, 1988.
- Ou, Y. C.; Ouyang, Y.; Lien, E. J. *J. Mol. Sci.* 1986, 4, 89.
- Tripos Associates, Inc. *Sybyl Version 6.2*; Tripos Associates, Inc.: St. Louis, MO, 1995.
- Tripos Associates, Inc. *CONCORD Version 3.2.1*; Tripos Associates, Inc.: St. Louis, MO, 1995.
- SAS Institute, Inc. In *SASSTAT User Guide, Release 6.03 Edition*; SAS Institute, Inc.: Cary, NC, 1988.

CI960024I

Use of Topostructural, Topochemical, and Geometric Parameters in the Prediction of Vapor Pressure: A Hierarchical QSAR Approach

Subhash C. Basak, Brian D. Gute, and Gregory D. Grunwald

Natural Resources Research Institute, University of Minnesota,
Duluth, Duluth, Minnesota 55811

Journal of
**Chemical
Information and
Computer Sciences®**

Reprinted from
Volume 37, Number 4, Pages 651-655

Use of Topostructural, Topochemical, and Geometric Parameters in the Prediction of Vapor Pressure: A Hierarchical QSAR Approach

Subhash C. Basak,* Brian D. Gute, and Gregory D. Grunwald

Natural Resources Research Institute, University of Minnesota, Duluth, Duluth, Minnesota 55811

Received December 31, 1996*

Numerous quantitative structure-activity relationships (QSARs) have been developed using topostructural, topochemical, and geometrical molecular descriptors. However, few systematic studies have been carried out on the relative effectiveness of these three classes of parameters in predicting properties. We have carried out a systematic analysis of the relative utility of the three types of structural descriptors in developing QSAR models for predicting vapor pressure at STP for a set of 476 diverse chemicals. The hierarchical technique has proven to be useful in illuminating the relationships of different types of molecular description information to physicochemical property and is a useful tool for limiting the number of independent variables in linear regression modeling to avoid the problems of chance correlations.

1. INTRODUCTION

A large number of quantitative structure-activity relationship (QSAR) studies have been reported in recent literature using theoretical molecular descriptors in predicting physicochemical, pharmacological, and toxicological properties of molecules.¹⁻¹⁵ Such descriptors comprise graph invariants, geometrical or 3-D parameters, and quantum chemical indices. One of the reasons for the current upsurge of interest is the fact that such descriptors can be derived algorithmically, i.e., can be computed for any molecule, real or hypothetical, using standard software. Both in pharmaceutical drug design and in risk assessment of chemicals, one has to evaluate potential biological effects of chemicals. Evaluation schemes based on property-property correlation paradigms are not very useful in practical situations, because, for most of the candidate structures, the experimental data necessary for proper evaluation are not available. This is especially true for the thousands of chemicals rapidly produced by methods of combinatoric chemistry¹⁶ as well as for the large number of chemicals present in the Toxic Substances Control Act (TSCA) Inventory.¹⁷

A large number of physicochemical and biological endpoints are necessary for estimating the ecotoxicological fate, transport, and effects of environmental pollutants.¹⁷⁻¹⁹ The vapor pressure of chemicals is important in determining the partitioning of chemicals among different phases once they are released in the environment. Many QSARs have been reported for predicting normal vapor pressure of chemicals. Such studies are usually carried out on small sets of congeneric chemicals. Also, many QSARs use experimental data as inputs in the model. Therefore, it becomes necessary to develop QSARs based on nonempirical parameters which can predict the vapor pressure for a heterogeneous collection of chemicals so that such models are generally applicable. With this end in mind, in the current paper we have carried out a QSAR study of 476 diverse chemicals using three types of nonempirical molecular descriptors.

2. MATERIALS AND METHODS

2.1. Normal Vapor Pressure Database. Measured values for a subset of the Toxic Substances Control Act (TSCA) Inventory¹⁷ were obtained from the ASTER (Assessment Tools for the Evaluation of Risk) database.²⁰ This subset consisted of a diverse set of chemicals where vapor pressure (p_{vap}) was measured at 25 °C and over a pressure range of approximately 3–10 000 mmHg. Due to the size of the dataset being used in this study, data for these chemicals will not be listed in this paper. An electronic copy of the data may be obtained by contacting the authors.

2.2. Computation of Topological Indices. The majority of the topological indices (TIs) used in this study have been calculated by the computer program POLLY 2.3.²¹ These indices include Wiener index,²² the molecular connectivity indices developed by Randić and Kier and Hall,^{1,23} information theoretic indices defined on distance matrices of graphs,^{24,25} and a set of parameters derived on the neighborhood complexity of vertices in hydrogen-filled molecular graphs.²⁶⁻²⁸ Balaban's *J* indices²⁹⁻³¹ were calculated using software developed by the authors.

van der Waal's volume (V_w)³²⁻³⁴ was calculated using Sybyl 6.2.³⁵ The 3-D Wiener numbers³⁶ were calculated by Sybyl using an SPL (Sybyl Programming Language) program developed by the authors. Calculation of 3-D Wiener numbers consists of the summation of the entries in the upper triangular submatrix of the topographic Euclidean distance matrix for a molecule. The 3-D coordinates for the atoms were determined using CONCORD 3.2.1.³⁷ Two variants of the 3-D Wiener number were calculated, ${}^{3D}W_H$ and ${}^{3D}W$, where hydrogen atoms are included and excluded from the computations, respectively.

Table 1 provides a complete listing of all of the topological and geometrical parameters which have been used in this study. The listing includes the symbols used to represent the parameters and brief definitions for each of the parameters.

Two additional parameters were used in modeling normal vapor pressure, HB_1 , and dipole moment (μ). HB_1 is a simple hydrogen bonding parameter calculated using a program developed by Basak,³⁸ which is based on the ideas

* All correspondence should be addressed to Dr. Subhash C. Basak, Natural Resources Research Institute, University of Minnesota, Duluth, 5013 Miller Trunk Highway, Duluth, MN 55811.

© Abstract published in *Advance ACS Abstracts*, June 1, 1997.

Table 1. Symbols and Definitions of Topological and Geometrical Parameters

I_D^W	information index for the magnitudes of distances between all possible pairs of vertices of a graph
I_D^W	mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
P	degree complexity
H^V	graph vertex complexity
H^D	graph distance complexity
IC	information content of the distance matrix partitioned by frequency of occurrences of distance h
I_{orb}	information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
O	order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
M_1	a Zagreb group parameter = sum of square of degree over all vertices
M_2	a Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
IC_r	mean information content or complexity of a graph based on the r^h ($r = 0-5$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	structural information content for r^h ($r = 0-5$) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	complementary information content for r^h ($r = 0-5$) order neighborhood of vertices in a hydrogen-filled graph
χ	path connectivity index of order $h = 0-6$
χ_c	cluster connectivity index of order $h = 3-6$
χ_{pc}	path-cluster connectivity index of order $h = 4-6$
χ_{ca}	chain connectivity index of order $h = 5, 6$
χ_b	bond path connectivity index of order $h = 0-6$
χ_{bc}	bond cluster connectivity index of order $h = 3-6$
χ_{bc}	bond chain connectivity index of order $h = 5, 6$
χ_{bpc}	bond path-cluster connectivity index of order $h = 4-6$
χ^V	valence path connectivity index of order $h = 0-6$
χ^C	valence cluster connectivity index of order $h = 3-6$
χ^V_{ca}	valence chain connectivity index of order $h = 5, 6$
χ^V_{pc}	valence path-cluster connectivity index of order $h = 4-6$
P_h	number of paths of length $h = 0-10$
J	Balaban's J index based on distance
J^B	Balaban's J index based on bond types
J^X	Balaban's J index based on relative electronegativities
J^r	Balaban's J index based on relative covalent radii
V_w	van der Waal's volume
$3Dw$	3-D Wiener number for the hydrogen-suppressed geometric distance matrix
$3DW_H$	3-D Wiener number for the hydrogen-filled geometric distance matrix

of Ou *et al.*³⁹ Dipole moment was calculated using Sybyl 6.2.³⁵

2.3. Data Reduction. The set of 92 TIs was partitioned into two distinct subsets: topostructural indices and topochemical indices. The distinction was made as follows: topostructural indices encode information about the adjacency and distances of atoms (vertices) in molecular structures (graphs) irrespective of the chemical nature of the atoms involved in the bonding or factors like hybridization states of atoms and number of core/valence electrons in individual atoms, while topochemical indices quantify information regarding the topology (connectivity of atoms) as well as specific chemical properties of the atoms comprising a molecule. Topochemical indices are derived from weighted molecular graphs where each vertex (atom) is properly weighted with selected chemical/physical properties. These subsets are shown in Table 2.

The partitioning of the indices left 38 topostructural indices and 54 topochemical indices. At this point no further data reduction is called for, since the ratio of the number of

Table 2. Classification of Parameters used in Modeling Normal Vapor Pressure [$\log_{10}(P_{\text{vap}})$]

topological	topochemical	geometric	other parameters
I_D^W	I_{orb}	V_w	HB_1
I_D^W	IC_0-IC_5	$3DW$	μ
W	SIC_0-SIC_5	$3DW_H$	
P	CIC_0-CIC_5		
H^V	$\chi_0^0-\chi_6^0$		
H^D	$\chi_3^0-\chi_6^0$		
IC	$\chi_5^0-\chi_6^0$		
O	$\chi_4^0-PC-\chi_6^0-PC$		
M_1	$\chi_4^0-\chi_6^0$		
M_2	$\chi_3^0-\chi_6^0$		
χ	$\chi_5^0-\chi_6^0$		
χ_c	$\chi_5^0-\chi_6^0$		
χ_{pc}	$\chi_5^0-PC-\chi_6^0-PC$		
χ_{ca}	$\chi_5^0-\chi_6^0$		
χ_{bc}	$\chi_5^0-\chi_6^0$		
χ_{bpc}	$\chi_5^0-\chi_6^0$		
χ^V	$\chi_5^0-\chi_6^0$		
χ^C	$\chi_5^0-\chi_6^0$		
χ^V_{ca}	$\chi_5^0-\chi_6^0$		
χ^V_{pc}	$\chi_5^0-\chi_6^0$		
P_h	J^B		
J	J^X		
J^B	J^r		
J^X			
J^r			
V_w			
$3DW$			
$3DW_H$			
J			

observations in the training set (342) to the total number of variables (92 maximum) falls well within the condition limits suggested by Topliss and Edwards⁴⁰ for reducing the probability of spurious correlations even at the more conservative $R^2 \geq 0.7$ level.

2.4. Statistical Analysis and Hierarchical QSAR. Initially, all TIs were transformed by the natural logarithm of the index plus one. This was done since the scale of some indices may be several orders of magnitude greater than that of other indices. The geometric parameters were transformed by the natural logarithm of the parameter.

Two regression procedures were used in developing the linear models. When the number of independent variables was high, typically greater than 25, a stepwise regression procedure was used to maximize the improvement of the explained variance (R^2). When the number of independent variables was smaller, all possible subsets regression was used. Models were then optimized to reduce problems of variance inflation and collinearity. Regression modeling was conducted using the REG procedure of the statistical package SAS.⁴¹

The vapor pressure data (P_{vap}) was split into a training set (342 compounds) and a test set (134 compounds), an approximately 75/25 split. Models were developed using the training set of chemicals and then used to predict the P_{vap} values of the test chemicals. Final models were then developed using the combined training and test set of chemicals.

Five sets of indices were used in model development. These sets were constructed as part of a hierarchical approach to QSAR modeling. The hierarchy begins with the simplest indices, the topostructural. After developing our initial model utilizing the topostructural indices, we increase the level of complexity. To the indices included in the best topostructural model, we add all of the topochemical indices and proceed to model P_{vap} using these parameters. Likewise, the indices included in the best model from this procedure are combined with the geometrical indices and modeling is conducted once again. In addition to this hierarchical approach, models were also constructed using the topochemical indices alone and the geometrical indices alone for purposes of comparison.

3. RESULTS

Stepwise regression analyses for $\log_{10}(P_{\text{vap}})$ of the training set of chemicals is summarized in Table 3. As shown in

Table 3. Summary of the Regression Results for the Training Set and the Prediction Results for the Test Set for the Hierarchical Analysis of $\log_{10}(p_{\text{vap}})$

parameter class	variables included	training set ($N = 342$)			test set ($N = 134$)	
		F	R^2	s	R^2	s
topostructural	$^1\chi$, $^6\chi_C$, P_9	104.6	48.1	0.56	57.9	0.46
topochemical	SIC_0 , SIC_2 , SIC_3 , CIC_0 , CIC_1 , $^3\chi^b_C$, $^1\chi^v$, $^5\chi^v$, $^3\chi^v_C$, J^Y	126.3	79.2	0.36	85.8	0.27
geometrical	^{3D}W , $^{3D}W_H$, V_W	168.9	51.8	0.53	62.2	0.44
topostructural + topochemical	$^1\chi$, P_9 , IC_1 , SIC_2 , CIC_1 , $^3\chi^b_C$, $^1\chi^v$, $^3\chi^v$, $^6\chi^v$, $^3\chi^v_C$, $^5\chi^v_{CH}$	112.5	80.4	0.35	84.7	0.28
all indices	H^V , SIC_1 , SIC_2 , CIC_0 , CIC_3 , $^6\chi_C$, $^1\chi^v$, $^3\chi^v$, $^6\chi^v_C$, P_6 , P_{10}	117.4	79.6	0.35	84.2	0.28
$tg + HB_1 + \mu$	$^1\chi$, P_3 , P_9 , IC_0 , $^3\chi^b_C$, $^1\chi^v$, $^3\chi^v$, $^5\chi^v_C$, HB_1	160.8	82.9	0.32	83.1	0.29

Table 3, the topostructural model using three parameters resulted in an explained variance (R^2) of 48.1% and a standard error (s) of 0.56. Addition of the topochemical parameters to the three topostructural parameters led to a significant increase in the effectiveness of the model. The resulting model used 12 parameters, two topostructural and ten topochemical. This model had an R^2 of 80.4% and s of 0.35. All subsets regression of the two topostructural and ten topochemical indices retained thus far and the three geometrical indices resulted in the selection of the same 12 parameter model, thus the geometrical indices did not contribute significantly to model development. Several other models were constructed for comparative purposes. Using topochemical indices only, a ten parameter model was developed which had an R^2 of 79.2% and s of 0.36. A geometrical model was developed which utilized all three geometrical indices and resulted in an R^2 of 51.8% and s of 0.53. Finally, two additional stepwise models were developed. One model simply used all indices for a comparison between a simple stepwise analysis of the data and the results of the hierarchical procedure. This resulted in an 11 parameter model with R^2 of 79.6% and s of 0.35. The second model added two new parameters, HB_1 and μ . We thought that it might be possible to improve our modeling by adding in some other nonempirical parameters which could be important to the determination of normal vapor pressure. We selected the parameters HB_1 and μ , since they would be important in intermolecular interactions which could have a dramatic effect on vapor pressure. To look at the addition of these parameters, we conducted a stepwise regression analysis using all topostructural, topochemical, and geometric indices so that we would be able to optimize our model, just as we had done with the previous models. The addition of these parameters led to the selection of a ten parameter model which included three topostructural indices, nine topochemical indices, and HB_1 . This was the best model yet, with an R^2 of 82.9% and s of 0.32.

Application of these six models to the test set of chemicals resulted in comparable R^2 and s; actually all models improved slightly on their predictions of the test set, and these values are also listed in Table 3. Based on these results, we decided that it was pointless to develop further models using only geometrical parameters. Also, based on the findings that the geometrical indices did not contribute significantly to any of the training models, they were dropped from the development of final models for the full set of 476 chemicals. However, even though the topostructural indices did not perform well in modeling vapor pressure by themselves, they will be used in model development since they did contribute significantly to most of the models.

Regression analyses of the combined set of 476 chemicals showed similar results for estimating $\log_{10}(p_{\text{vap}})$ as analysis

of the training set. Using only the topostructural indices, stepwise regression analysis resulted in a five parameter model to estimate vapor pressure:

$$\log_{10}(p_{\text{vap}}) = 4.88 + 0.20(O) - 2.56(^1\chi) + 0.49(^4\chi_C) + 0.79(^6\chi_C) + 0.98(P_{10}) \quad (1)$$

$$n = 476, R^2 = 51.5\%, s = 0.53, F = 99.7$$

Stepwise regression using the five topostructural parameters and all topochemical parameters resulted in the selection of the following seven parameter model:

$$\log_{10}(p_{\text{vap}}) = 8.44 - 1.77(^1\chi) + 1.25(P_{10}) - 5.69(IC_1) + 3.91(IC_0) - 1.24(IC_5) + 1.41(^3\chi^b_C) - 1.70(^1\chi^v) \quad (2)$$

$$n = 476, R^2 = 79.3\%, s = 0.34, F = 224.0$$

Only two of the topostructural indices used in eq 1 were retained by the stepwise regression procedure used to produce eq 2: $^1\chi$ and P_{10} . The improvement in R^2 was significant, increasing from 51.5% for eq 1 to 79.3% for eq 2. Also, the model error decreased significantly, dropping by 0.19 logarithmic units. Since we have dropped the geometrical indices, this becomes our final hierarchical model.

The stepwise regression analysis of only topochemical parameters resulted in a 12 parameter model:

$$\log_{10}(p_{\text{vap}}) = 6.65 - 3.44(IC_0) - 1.33(IC_5) + 3.47(SIC_2) + 0.87(CIC_1) - 0.48(^4\chi^b) + 1.44(^3\chi^b_C) - 1.00(^1\chi^v) - 0.41(^3\chi^v) - 0.70(^5\chi^v) - 1.08(^3\chi^v_C) + 1.42(^6\chi^v_{CH}) - 1.23(J^Y) \quad (3)$$

$$n = 476, R^2 = 75.8\%, s = 0.38, F = 120.5$$

This model which is inferior to the topostructural + topochemical model (eq 2), because its variance explained is lower and, more importantly, it requires more independent variables (parameters) to achieve this explanation of variance.

Stepwise regression of all indices resulted in the selection of an 11 parameter model. This approach selected three topostructural indices and eight topochemical indices to arrive at the following model:

$$\log_{10}(p_{\text{vap}}) = 7.85 - 2.56(H^V) + 1.17(^6\chi_C) - 5.01(IC_1) + 3.65(IC_2) - 0.99(IC_5) + 0.51(CIC_1) - 1.54(^1\chi^v) - 0.36(^3\chi^v) - 0.36(^4\chi^v) - 1.40(^6\chi^v_C) \quad (4)$$

$$n = 476, R^2 = 80.4\%, s = 0.33, F = 173.4$$

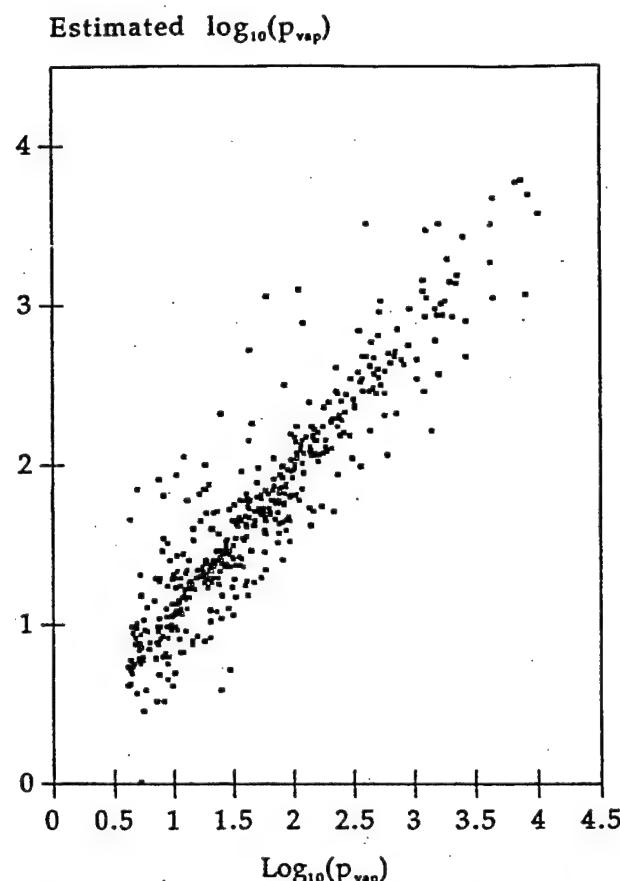


Figure 1. Scatterplot of observed $\log_{10}(p_{\text{vap}})$ vs estimated $\log_{10}(p_{\text{vap}})$ using eq 5 for 476 diverse compounds.

While eq 4 shows some slight improvements over eq 2, the hierarchical model, eq 2 is preferred since it is a simpler model using seven indices instead of 11 and based on a comparison of F values it is a more robust model than that in eq 4.

Finally, we conducted the stepwise regression modeling using all topostructural and topochemical indices with HB_1 and μ for the complete set of 476 chemicals. The resulting ten parameter model used three topostructural indices, six topochemical indices, and HB_1 :

$$\begin{aligned} \log_{10}(p_{\text{vap}}) = & 9.67 - 3.66(^1\chi) + 0.35(\text{P}_3) + 0.74(\text{P}_9) - \\ & 1.78(\text{IC}_0) - 3.33(\text{SIC}_1) - 0.81(\text{CIC}_2) + 2.05(^2\chi^b) - \\ & 1.73(^2\chi^v) - 0.79(^3\chi^v) - 0.29(\text{HB}_1) \quad (5) \end{aligned}$$

$$n = 476, R^2 = 84.3\%, s = 0.29, F = 249.5$$

Equation 5 shows marked improvement over eq 2, justifying the addition of indices to the model. Also, it meets the criteria on which eq 4 was judged to be lacking. Overall, there is an improvement in variance explained of 5%, with a comparable decrease in standard deviation. A scatter plot of observed $\log_{10}(p_{\text{vap}})$ versus estimated $\log_{10}(p_{\text{vap}})$ using eq 5 is presented in Figure 1.

4. DISCUSSION

The purpose of this paper was 2-fold: (a) to study the utility of algorithmically-derived molecular descriptors in developing QSAR models for predicting the vapor pressure of chemicals from structure and b) to investigate the relative

Table 4. Summary of the Chemical Class Composition of the Normal Vapor Pressure Dataset

compd classification	no. of compds	pure	substituted
total normal vapor pressure dataset	476		
hydrocarbons ^a	253		
non-hydrocarbons ^a	223		
nitro compounds	4	3	1
amines	20	17	3
nitriles	7	6	1
ketones	7	7	0
halogens	100	95	5
anhydrides	1	1	0
esters	18	16	2
carboxylic acids	2	2	0
alcohols	10	6	4
sulfides	39	38	1
thiols	4	4	0
imines	2	2	0
epoxides	1	1	0
aromatic compounds ^b	15	10	4
fused-ring compounds ^c	1	1	0

^a The non-hydrocarbons are further broken down into the following groups. ^b The 15 aromatic compounds are a mixture of 11 aromatic hydrocarbons and four aromatic halides. ^c The only fused-ring compound was a polycyclic aromatic hydrocarbon.

roles of topostructural, topochemical, and geometrical indices in the estimation of standard vapor pressure.

Results described in this paper (eqs 1–5) show that nonempirical parameters derived predominantly from graph theoretic models of molecules can estimate normal vapor pressure of diverse chemicals reasonably well. The explained variance of data ($R^2 = 84.3\%$) is excellent in view of the fact that the database of chemicals analyzed in this paper is very diverse (see Table 4). It should be mentioned that most published QSAR models for the estimation of vapor pressure have dealt with much smaller data sets with limited structural variety.^{42,43}

The relative effectiveness of topostructural, topochemical, and geometrical indices in predicting normal vapor pressure of chemicals is evident from the result presented above. Equation 1 explains over 51% of variance in the data. All parameters used to derive eq 1 are topostructural, *i.e.*, they are parameters which encode information about the adjacency and distance of vertices in skeletal molecular graphs without quantifying any explicit information about such chemical aspects like bond order, electronic character of atoms, etc. Yet, the high explained variance of the property indicates that adjacency and distance in chemical graphs, being general descriptors of molecular size, shape, and branching, are important in predicting properties. This may explain the success of parameters like simple connectivity indices in estimating many diverse properties.¹

Equation 3 is derived only from topochemical indices. The explained variance of vapor pressure (75.8%) shows that topochemical parameters, as a class, explain a larger fraction of the variance as compared to models derived from only topostructural indices (eq 1). Geometrical parameters were dropped from the set of descriptors after their limited success in prediction for the training and test sets. This is in line with our earlier studies with normal boiling point and hydrophobicity, where it was reported that the addition of geometrical indices could not significantly improve the predictive power of QSAR models derived from a combined set of topostructural and topochemical parameters.¹⁵ It would

be interesting to see whether this pattern holds good for other properties as well. Finally, the addition of the simple nonempirical parameter, HB₁, which contains information relevant to intermolecular interactions further improves our ability to estimate normal vapor pressure resulting in an explained variance of 84.3% (eq 5).

ACKNOWLEDGMENT

This is contribution number 209 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported in part by grants F49620-94-1-0401 and F49620-96-1-0330 from the United States Air Force, a grant from Exxon Corporation and the Structure-Activity Relationship Consortium (SARCON) of the Natural Resources Research Institute of the University of Minnesota. The authors would like to dedicate this paper to Professor Milan Randić in appreciation of his contributions in chemical information, quantitative structure-activity relationships, and chemical graph theory.

REFERENCES AND NOTES

- (1) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Research Studies Press: Chichester, England, 1986.
- (2) Basak, S. C. Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: A QSAR Approach. *Med. Sci. Res.* 1987, 15, 605-609.
- (3) Balaban, A. T.; Basak, S. C.; Colburn, T.; Grunwald, G. Correlation Between Structure and Normal Boiling Points of Haloalkanes C₁-C₄ Using Neural Networks. *J. Chem. Inf. Comput. Sci.* 1994, 34, 1118-1121.
- (4) Basak, S. C. A Nonempirical Approach to Predicting Molecular Properties Using Graph-Theoretic Invariants. In *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*; Karcher, W., Devillers, J., Eds.; Kluwer Academic Publishers: Dordrecht/Boston/London, 1990; pp 83-103.
- (5) Basak, S. C.; Bertelsen, S.; Grunwald, G. Application of Graph Theoretical Parameters in Quantifying Molecular Similarity and Structure-Activity Studies. *J. Chem. Inf. Comput. Sci.* 1994, 34, 270-276.
- (6) Basak, S. C.; Bertelsen, S.; Grunwald, G. D. Use of Graph Theoretic Parameters in Risk Assessment of Chemicals. *Toxicol. Lett.* 1995, 79, 239-250.
- (7) Basak, S. C.; Grunwald, G. D. Use of Topological Space and Property Space in Selecting Structural Analogs. *Mathematical Modelling and Scientific Computing*. In press.
- (8) Basak, S. C.; Grunwald, G. D. Molecular Similarity and Risk Assessment: Analog Selection and Property Estimation Using Graph Invariants. *SAR and QSAR in Environ. Res.* 1994, 2, 289-307.
- (9) Basak, S. C.; Grunwald, G. D. Estimation of Lipophilicity From Molecular Structural Similarity. *New J. Chem.* 1995, 19, 231-237.
- (10) Basak, S. C.; Grunwald, G. D.; Niemi, G. J. Use of Graph-Theoretic and Geometrical Molecular Descriptors in Structure-Activity Relationships. In *From Chemical Topology To Three-Dimensional Geometry*; Balaban, A. T., Ed.; Plenum Press: New York, 1997; pp 73-116.
- (11) Basak, S. C.; Gute, B. D. Use of Graph Theoretic Parameters in Predicting Inhibition of Microsomal Hydroxylation of Anilines by Alcohols: A Molecular Similarity Approach. In *Proceedings of the International Congress on Hazardous Waste: Impact on Human and Ecological Health*; Johnson, B. L., Xintaras, C., Andrews, J. S., Jr. Eds.; Princeton Scientific Publishing Co., Inc.: Princeton, NJ, 1997; pp 492-504.
- (12) Basak, S. C.; Gute, B. D.; Drewes, L. R. Predicting Blood-Brain Transport of Drugs: A Computational Approach. *Pharm. Res.* 1996, 13, 775-778.
- (13) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Development and Applications of Molecular Similarity Methods Using Nonempirical Parameters. *Math. Modelling Sci. Computing* In press.
- (14) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Estimation of Normal Boiling Points of Haloalkanes Using Molecular Similarity. *Croat. Chem. Acta.* 1996, 69, 1159-1173.
- (15) Basak, S. C.; Gute, B. D.; Grunwald, G. D. A Comparative Study of Topological and Geometrical Parameters in Estimating Normal Boiling Point and Octanol/Water Partition Coefficient. *J. Chem. Inf. Comput. Sci.* 1996, 36, 1054-1060.
- (16) Martin, Y. C. Opportunities for Computational Chemists Afforded by the New Strategies in Drug Discovery: An Opinion. *Network Science* 1996.
- (17) Auer, C. M.; Nabholz, J. V.; Baetcke, K. P. Mode of Action and the Assessment of Chemical Hazards in the Presence of Limited Data: Use of Structure-Activity Relationships (SAR) Under TSCA, Section 5. *Environ. Health Perspect.* 1990, 87, 183-197.
- (18) NRC. *Toxicity Testing: Strategies to Determine Needs and Priorities*; National Academy Press: Washington, DC, 1984; p 84.
- (19) Basak, S. C.; Niemi, G. J.; Veith, G. D. Predicting Properties of Molecules Using Graph Invariants. *J. Math. Chem.* 1991, 7, 243-272.
- (20) Russom, C. L.; Anderson, E. B.; Greenwood, B. E.; Pilli, A. ASTER: An Integration of the AQUIRE Data Base and the QSAR System for Use in Ecological Risk Assessments. *Sci. Total Environ.* 1991, 109/110, 667-670.
- (21) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. *POLLY Version 2.3*; Copyright of the University of Minnesota, 1988.
- (22) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* 1947, 69, 17-20.
- (23) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* 1975, 97, 6609-6615.
- (24) Raychaudhury, C.; Ray, S. K.; Ghosh, J. J.; Roy, A. B.; Basak, S. C. Discrimination of Isomeric Structures Using Information Theoretic Topological Indices. *J. Comput. Chem.* 1984, 5, 581-588.
- (25) Bonchev, D.; Trinajstić, N. Information Theory, Distance Matrix and Molecular Branching. *J. Chem. Phys.* 1977, 67, 4517-4533.
- (26) Basak, S. C.; Magnuson, V. R. Molecular Topology and Narcosis: A Quantitative Structure-Activity Relationship (QSAR) Study of Alcohols Using Complementary Information Content (CIC). *Arzneim. Forsch.* 1983, 33, 501-503.
- (27) Basak, S. C.; Roy, A. B.; Ghosh, J. J. In *Proceedings of the Second International Conference on Mathematical Modelling*; Avula, X. J. R., Bellman, R., Luke, Y. L., Rigler, A. K., Eds.; University of Missouri-Rolla: Rolla, MO, 1980; p 745.
- (28) Roy, A. B.; Basak, S. C.; Harriss, D. K.; Magnuson, V. R. Neighborhood Complexities and Symmetry of Chemical Graphs and Their Biological Applications. In *Mathematical Modelling in Science and Technology*; Avula, X. J. R., Kalman, R. E., Liapis, A. L., Rodin, E. Y., Eds.; Pergamon: New York, 1984; p 745.
- (29) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* 1982, 89, 399-404.
- (30) Balaban, A. T. Topological Indices Based on Topological Distances in Molecular Graphs. *Pure Appl. Chem.* 1983, 55, 199-206.
- (31) Balaban, A. T. Chemical Graphs. Part 48. Topological Index J for Heteroatom-Containing Molecules Taking into Account Periodicities of Element Properties. *Math. Chem. (MATCH)* 1986, 21, 115-122.
- (32) Bondi, A. van der Waal's Volumes and Radii. *J. Phys. Chem.* 1964, 68, 441-451.
- (33) Moriguchi, I.; Kanada, Y. Use of van der Waal's Volume in Structure-Activity Studies. *Chem. Pharm. Bull.* 1977, 25, 926-935.
- (34) Moriguchi, I.; Kanada, Y.; Komatsu, K. van der Waal's Volume and the Related Parameters for Hydrophobicity in Structure-Activity Studies. *Chem. Pharm. Bull.* 1976, 24, 1799-1806.
- (35) Tripos Associates, Inc. *SYBYL Version 6.2*; Tripos Associates, Inc.: St. Louis, MO, 1994.
- (36) Bogdanov, B.; Nikolić, S.; Trinajstić, N. On the Three-Dimensional Wiener Number. *J. Math. Chem.* 1989, 3, 299-309.
- (37) Tripos Associates, Inc. *CONCORD Version 3.2.7*; Tripos Associates, Inc.: St. Louis, MO, 1995.
- (38) Basak, S. C. *H-Bond*; Copyright of the University of Minnesota, 1988.
- (39) Ou, Y. C.; Ouyang, Y.; Lien, E. J. *J. Mol. Sci.* 1986, 4, 89.
- (40) Topliss, J. G.; Edwards, R. P. Chance Factor in Studies of Quantitative Structure-Activity Relationships. *J. Med. Chem.* 1979, 22, 1238-1244.
- (41) SAS Institute Inc. In *SAS/STAT User's Guide, Release 6.03 Edition*. SAS Institute Inc.: Cary, NC, 1988; Chapters 28 and 34, pp 773-875, 949-965.
- (42) Drefahl, A.; Reinhard, M. *Handbook for Estimating Physico-Chemical Properties of Organic Compounds*; Stanford University Bookstore, Stanford, CA, 1995.
- (43) Lyman, W. J.; Reehl, W. F.; Rosenblatt, D. H. *Handbook of Chemical Property Estimation Methods: Environmental Behavior of Organic Compounds*; McGraw-Hill Book Company: New York, 1982.

PREDICTING ACUTE TOXICITY (LC_{50}) OF BENZENE DERIVATIVES USING
THEORETICAL MOLECULAR DESCRIPTORS: A HIERARCHICAL QSAR
APPROACH

Brian D. Gute
and
Subhash C. Basak*

Natural Resources Research Institute, University of Minnesota,
5013 Miller Trunk Highway
Duluth, MN 55811, USA

SAR and QSAR in Environmental Research, in press, 1997.

*Author to whom all correspondence should be addressed

ABSTRACT

Four classes of theoretical structural parameters, viz., topostructural, topochemical, geometrical and quantum chemical descriptors, have been used in the development of quantitative structure-activity relationship (QSAR) models for a set of sixty-nine benzene derivatives. None of the individual classes of parameters was very effective in predicting toxicity. A hierarchical approach was followed in using a combination of the four classes of indices in QSAR model development. The results show that the hierarchical QSAR approach using the algorithmically derived molecular descriptors can estimate the LC_{50} values of the benzene derivatives reasonably well.

KEYWORDS

hierarchical QSAR; topological indices; geometrical indices; quantum chemical parameters; aquatic toxicity; benzene derivatives

INTRODUCTION

Today's toxicologist is faced with a myriad of unknowns. In 1996 approximately 1.26 million new chemicals were registered with the Chemical Abstract Service (CAS), bringing the total number of registered chemicals to around 15.8 million [1]. With such a large number of chemicals being registered yearly, it is impossible to test all of them exhaustively for their effects on the environment and human health. Chemicals can only be evaluated as they are called into question, and for many of these compounds there will be little or no test data available. Therefore, when the issue of hazard assessment comes up, it becomes difficult at best to provide any useful suggestions or analyses for many of the registered chemicals, including some which are in commerce today. To complete the battery of tests necessary for the proper hazard assessment of a single compound is an extremely costly procedure and there is simply not enough time or money to complete these test batteries for all compounds which are registered today [2]. As a result, when we need to evaluate the human health or ecological hazards posed by a chemical it becomes ever more important that we have accurate methods for estimating the physicochemical and biological properties of molecules.

Quantitative structure-activity relationships (QSARs) have come into widespread use for the prediction of various molecular properties and biological responses. Traditional QSARs use empirical properties; e.g., boiling point, melting point, octanol-water partition coefficient; or empirically derived parameters; e.g., linear free energy related (LFER) and linear solvation energy related (LSER) parameters; for the

prediction of other endpoints [3-8]. However, due to the scarcity of available data for the majority of chemicals that need to be evaluated for ecotoxicological risk assessment, these physicochemical properties necessary for traditional QSAR model development may not be known. When this is the case, it is imperative that we have methods that make use of nonempirical parameters. One of the fundamental principles of biochemistry is that activity is dictated by structure [9]. Following this principle, one can use theoretical molecular descriptors which quantify structural aspects of the molecular structure [10-27]. These theoretical descriptors can be generated directly from the molecular structure alone, without any input of experimental data.

Topological indices (TIs) are numerical graph invariants that quantify certain aspects of molecular structure. TIs are sensitive to such structural features as size, shape, bond order, branching, and neighborhood patterns of atoms in molecules. They can be derived from simple linear graphs, multigraphs, weighted graphs, and weighted pseudographs. TIs derived from these different classes of graphs will encode different types of information about molecular architecture. The different classes of TIs provide us with nonempirical, quantitative descriptors that can be used in place of experimentally derived descriptors in QSARs for the prediction of properties.

Our recent studies have focused on the role of different classes of theoretical descriptors of increasing levels of complexity and their utility in QSAR [28-31]. This takes the form of a hierarchical approach which examines the relative contributions of parameters of gradually increasing complexity; e.g., structural, chemical, shape, and quantum chemical descriptors; in estimating physicochemical and biological properties.

In this paper we have reported the utility of this hierarchical approach in modeling the acute aquatic toxicity (LC_{50}) of a congeneric set of sixty-nine benzene derivatives.

THEORETICAL METHODS

Database

Acute aquatic toxicity [-log(LC_{50})] in fathead minnow (*Pimephales promelas*) data was taken from the work of Hall, Kier and Phipps [32]. Their data was compiled from eight other sources, as well as some original work which was conducted at the U.S. Environmental Protection Agency (USEPA) Environmental Research Laboratory in Duluth, Minnesota. The complete set of fathead minnow data included 69 benzene derivatives. According to the authors, the set of benzene derivatives were tested using methodologies which were comparable to their 96-hour fathead minnow toxicity test system. The derivatives chosen for this study have seven different substituent groups that are all present in at least six of the molecules. These groups consist of chloro, bromo, nitro, methyl, methoxyl, hydroxyl, and amino substituents (Table I).

Computation of Indices

Four distinct sets of theoretical descriptors have been used in this study. These sets include topostructural, topochemical, geometric, and quantum chemical indices. The

topostructural and topochemical indices fall into the category normally grouped together as topological indices. The geometrical indices are three-dimensional Wiener number for hydrogen-filled molecular structure, hydrogen-suppressed molecular structure, and van der Waals volume.

Topostructural indices (TSIs) are topological indices which only encode information about the adjacency and distances of atoms (vertices) in molecular structures (graphs), irrespective of the chemical nature of the atoms involved in bonding or factors such as hybridization states and the number of core/valence electrons in individual atoms. Topochemical indices (TCIs) are parameters that quantify information regarding the topology (connectivity of atoms), as well as specific chemical properties of the atoms comprising a molecule. These indices are derived from weighted molecular graphs where each vertex (atom) or edge (bond) is properly weighted with selected chemical or physical property information. The sets of topostructural and topochemical indices are shown in Table II.

Topological Indices

The 102 topological indices used in this study, both the topostructural and the topochemical, have been calculated by POLLY 2.3 [33] and software developed by the authors. These indices include Wiener index [34], connectivity indices developed by Randić [35] and higher order connectivity indices formulated by Kier and Hall [36], bonding connectivity indices defined by Basak *et al.* [37], a set of information theoretic

indices defined on the distance matrices of simple molecular graphs [38,39] and neighborhood complexity indices of hydrogen-filled molecular graphs [40,41], and Balaban's *J* indices [42-44]. Table III provides the list of the topostructural, topochemical, and geometrical indices included in this study.

Geometrical Indices

Van der Waals volume, V_w [45-47], was calculated using *Sybyl 6.1* from Tripos Associates, Inc [48]. The 3-D Wiener numbers were calculated by *Sybyl* using an SPL (Sybyl Programming Language) program developed in our lab [49]. Calculation of 3-D Wiener numbers consists of the sum entries in the upper triangular submatrix of the topographic Euclidean distance matrix for a molecule. The 3-D coordinates for the atoms were determined using *CONCORD 3.0.1* [50]. Two variants of the 3-D Wiener number were calculated: ${}^{3D}W_H$ and ${}^{3D}W$. For ${}^{3D}W_H$, hydrogen atoms are included in the computations and for ${}^{3D}W$, hydrogen atoms are excluded from the computations.

Quantum Chemical Parameters

The following quantum chemical parameters were calculated using the Austin Model version one (AM1) semi-empirical Hamiltonian: energy of the highest occupied molecular orbital (E_{HOMO}), energy of the second highest occupied molecular orbital (E_{HOMO1}), energy of the lowest unoccupied molecular orbital (E_{LUMO}), energy of the second

lowest unoccupied molecular orbital (E_{LUMO1}), heat of formation (ΔH_f), and dipole moment (μ). These parameters were calculated using *MOPAC 6.00* in the *SYBYL* interface [51].

Data Reduction

Initially, all topological indices were transformed by the natural logarithm of the index plus one. This was done to scale the indices, since some may be several orders of magnitude greater than others, while other indices may equal zero. The geometric indices were transformed by the natural logarithm of the index for consistency, the addition of one was unnecessary.

The set of eighty-one topological indices was then partitioned into two distinct sets, the topostructural indices (thirty-four) and the topochemical indices (forty-seven). To further reduce the number of independent variables for model construction, the sets of topostructural and topochemical indices were further divided into subsets, or clusters, based on the correlation matrix using the SAS procedure VARCLUS [52]. This procedure divides the set of indices into disjoint clusters, such that each cluster is essentially unidimensional.

From each cluster we selected the index most correlated with the cluster, as well as any indices which were poorly correlated with their cluster ($R^2 < 0.70$). These indices were then used in the modeling of the acute aquatic toxicity of benzene derivatives in fathead minnow. The variable clustering and selection of indices was performed independently for both the topostructural and topochemical indices. This procedure

resulted in a set of five topostructural indices and a set of nine topochemical indices.

Reducing the number of independent variables is critical when attempting to model small datasets. The smaller the dataset is, the greater the chance of spurious error when using a large number of independent variables (descriptors). Topliss and Edwards have studied this issue of chance correlations [53]. For a set with about seventy dependent variables (observations), to keep the probability of chance correlations less than 0.01, we can use at most forty independent variables. This number is dependent on the actual correlation achieved in the modeling process, with a high correlation we have a better chance of using more variables with the same limited probability of chance correlations. In this study we are well below the cut-off of forty. In fact, the total number of descriptors which will be used for model construction and estimation is twenty-three, well within the bounds of the Topliss and Edwards criteria [53].

Statistical Analysis and Hierarchical QSAR

Regression modeling was accomplished using the SAS procedure REG on seven distinct sets of indices. These sets were constructed as part of a hierarchical approach to QSAR model development. The hierarchy begins with the simplest parameters, the TSIs. After using the TSIs to model the activity, the next level of complexity is added. To the indices included in the best TSI model, we add all of the TCIs and proceed to model the activity using these parameters. Likewise, the indices included in the best

model from this procedure are combined with the indices from the next level, the geometrical indices and modeling is conducted once again. Finally, the best model utilizing TSIs, TCIs and geometrical indices is combined with the quantum chemical parameters. The regression analysis results in the final selection of indices for each of the models. The remaining three models which use TCIs, geometric, and quantum chemical parameters independently serve as a means of validating the utility of the hierarchical approach and the need for varying types of theoretical descriptors.

RESULTS

The variable clustering of the topostructural indices resulted in the retention of five indices: M_1 , \overline{IC} , O , P_8 , P_9 . All-possible subsets regression resulted in the selection of a four-parameter model to estimate $-\log(LC_{50})$ with an explained variance (R^2) of 45.3% and a standard error (s) of 0.58. While this is an unsatisfactory model, the indices will still be retained and combined with the topochemical indices in the second step of model development. Table IV lists the indices used in each of the models.

The second step of the hierarchical method combined the four indices used in the first tier model with the nine topochemical indices selected in the variable clustering procedure: SIC_0 , SIC_1 , SIC_4 , CIC_0 , $^2\chi^b$, $^5\chi^b$, $^5\chi^v$, $^6\chi^v$, J^X . Again, all-possible subsets regression was conducted resulting in a four-parameter model with an explained variance (R^2) of 78.3% and a standard error (s) of 0.36. While this model retained two parameters from the topostructural model, it is evident that the addition of two

topochemical indices made a significant contribution to the effectiveness of our model.

The four indices from the second tier model were then combined with the three geometric parameters: ${}^{3D}W_H$, ${}^{3D}W$, V_w . The resulting model from this procedure retained four indices, replacing the topochemical index CIC_0 with the geometric parameter ${}^{3D}W_H$. This model had an explained variance (R^2) of 79.2% and a standard error (s) of 0.36.

The final step in the hierarchical method combined the four parameters from the third tier model with the quantum chemical (AM1) parameters: E_{HOMO} , E_{HOMO1} , E_{LUMO} , E_{LUMO1} , ΔH_f , μ . This set of ten indices led to a seven-parameter model with an explained variance (R^2) of 86.3% and a standard error (s) of 0.30. This model retained all of the indices from the third model and added three quantum chemical parameters.

Three other models were constructed for the purpose of comparison. These include a five-parameter topochemical model, a three parameter geometric model, and a four-parameter quantum chemical model. The indices used in these models and the results of the models can be found in Table IV.

DISCUSSION

The goal of this paper was to investigate the utility of hierarchical QSAR using algorithmically derived molecular descriptors in predicting LC_{50} values for a set of sixty-nine benzene derivaives. To this end, we used four classes of parameters, viz., topostructural descriptors, topochemical indices, geometrical descriptors and semi-empirical quantum chemical indices.

It is clear from the results described in Table IV that none of the individual classes of parameters correlate well with acute aquatic toxicity. The TSIs, the simplest of the four classes of parameters, explained about 45% of the variance in toxicity. The inclusion of topochemical indices in the set of independent variables made substantial improvement in the predictive capacity of the QSAR models. This is understandable since the benzene derivatives analyzed in this paper comprise a fairly congeneric set, and while the number and size of substituents may be important, the chemical nature of the substituents also plays an important role in determining the overall toxicity of the molecule. This is shown by the dramatic increase in predictive power between equations 1 and 2. Equation 2 replaces two TSI descriptors with two TCI indices that are sensitive to the atom types in all zero-order neighborhoods. The addition of this basic chemical information results in an improvement in the model. A similar conclusion is borne out from the QSAR analysis of the same set of benzene derivatives reported by Hall *et al.* where they found that the chemical nature of the substituent is important in determining toxicity [32].

In the next tier, equation 3 replaces one of the information content indices with the three-dimensional Wiener number, a descriptor that characterizes the three-dimensional aspects of molecular shape and size. This leads to refinement of the model developed in equation 2. Finally, the addition of the quantum chemical parameters; energy of the second lowest unoccupied molecular orbital , heat of formation, and dipole moment; leads to a marked improvement in the predictive power of the model (equation 4).

As can be seen from equations 1 and 5-7 (Table IV), none of the four classes of indices do very well individually. The hierarchical QSAR approach using four classes of parameters resulted in acceptable predictive models (equation 4). We may conclude from the results presented in this paper that each of the four classes of theoretical descriptors that were used are necessary for the development of good QSARs for the acute aquatic toxicity of benzene derivatives in fathead minnow.

ACKNOWLEDGMENTS

This is contribution number 213 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported in part by grants F49620-94-1-0401 and F49620-96-1-0330 from the United States Air Force, a grant from Exxon Corporation and the Structure-activity Relationship Consortium (SARCON) of the Natural Resources Research Institute of the University of Minnesota. The authors would like to extend their thanks to Greg Grunwald for technical support.

REFERENCES

1. Personal communication with W. Fisanick, Feb. 20, 1997.
2. Menzel, D.B. (1995). Extrapolating the future: research trends in modeling. *Toxicol. Lett.* **79**, 299-303.
3. Hansch, C. and Leo, A. (1995). *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*. American Chemical Society, Washington, D.C., p. 557.
4. Dearden, J.C. (1990). Physico-chemical descriptors, in *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology* (W. Karcher and J. Devillers, Eds.). Kluwer Academic Publishers, Dordrecht, pp. 25-59.
5. Lipnick, R.L. (1990). Narcosis: Fundamental and baseline toxicity mechanism for nonelectrolyte organic chemicals, in *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology* (W. Karcher and J. Devillers, Eds.). Kluwer Academic Publishers, Dordrecht, pp. 281-293.
6. Van de Waterbeemd, H. (1995). Discriminant analysis for activity prediction, in *Chemometric Methods in Molecular Design* (H. Van de Waterbeemd, Ed.). VCH Publishers, Inc., New York, pp. 283-294.
7. Kamlet, M.J., Abboud, J.-L.M., and Taft, R.W. (1977). Solvatochromic comparison method. 6. π^* scale of solvent polarities. *J. Am. Chem. Soc.* **99**, 6027-6038.
8. Kamlet, M.J., Abboud, J.-L.M., Abraham, M.H., and Taft, R.W. (1983). Linear solvation energy relationships. 23. A comprehensive collection of the solvatochromic parameters, π^* , α and β , and some methods for simplifying the generalized solvatochromic equation. *J. Org. Chem.* **48**, 2877-2887.
9. Hansch, C. (1976). On the structure of medicinal chemistry. *J. Med. Chem.* **19**, 1-6.
10. Randic, M. (1980). A graph theoretical approach to structure-property and structure-activity correlations. *Theoret. Chim. Acta (Berl.)* **58**, 45-68.
11. Randic, M. (1984). Nonempirical approach to structure-activity studies. *Int. J. Quant. Chem.* **11**, 137-153.
12. Randic, M. (1995). Molecular topographic indices. *J. Chem. Inf. Comput. Sci.* **35**, 140-147.
13. Sabljic, A. and Trinajstic, N. (1981). Quantitative structure-activity relationships: the role of topological indices. *Acta Pharm. Jugosl.* **31**, 189-214.
14. Basak, S.C. (1987). Use of molecular complexity indices in predictive pharmacology and toxicology: a QSAR approach. *Med. Sci. Res.* **15**, 605-609.

15. Balaban, A.T., Bertelsen, S., and Basak, S.C. (1994). New centric topological indexes for acyclic molecules (trees) and substituents (rooted trees), and coding of rooted trees. *MATCH* **30**, 55-72.
16. Basak, S.C. and Grunwald, G.D. (1995). Estimation of lipophilicity from molecular structural similarity. *New J. Chem.* **19**, 231-237.
17. Diudea, M.V., Horvath, D., and Graovac, A. (1995). Molecular topology. 15. 3D distance matrices and related topological indices. *J. Chem. Inf. Comput. Sci.* **35**, 129-135.
18. Estrada, E. (1995). Three-dimensional molecular descriptors based on electron charge density weighted graphs. *J. Chem. Inf. Comput. Sci.* **35**, 708-713.
19. Voelkel, A. (1994). Structural descriptors in organic chemistry - new topological parameter based on electrotopological state of graph vertices. *Computers Chem.* **18**, 1-4.
20. Basak, S.C., Gute, B.D., and Grunwald, G.D. (1996). Estimation of normal boiling points of haloalkanes using molecular similarity. *Croat. Chem. Acta* **69**, 1159-1173.
21. Basak, S.C., Gute, B.D., and Drewes, L.R. (1996). Predicting blood-brain transport of drugs: a computational approach. *Pharm. Res.* **13**, 775-778.
22. Basak, S.C., Gute, B.D., and Grunwald, G.D. (1997). Development and application of molecular similarity methods: using nonempirical parameters. *Mathl. Model. And Sci. Comput.*, in press.
23. Basak, S.C. and Gute, B.D. (1997). Use of graph-theoretic parameters in predicting inhibition of microsomal *p*-hydroxylation of aniline by alcohols: a molecular similarity approach, in *Proceedings of the 2nd international Congress on Hazardous Waste: Impact on Human and Ecological Health* (B.L. Johnson, C. Xintaras, and J.S. Andrews, Jr., Eds.). Princeton Scientific Publishing Co., Inc., New Jersey, pp. 492-504.
24. Basak, S.C., Gute, B.D., and Ghatak, S. (1997). Prediction of complement-inhibitory activity of benzamidines using topological and geometric parameters. *J. Chem. Inf. Comput. Sci.*, submitted.
25. Famini, G.R., Penski, C.A., and Wilson, L.Y. (1992). Using theoretical descriptors in quantitative structure activity relationships: some physicochemical properties. *J. Phys. Org. Chem.* **5**, 395-408.
26. Cramer, C.J., Famini, G.R., and Lowrey, A.H. (1993). Use of calculated quantum chemical properties as surrogates for solvatochromic parameters in structure-activity relationships. *Acc. Chem. Res.* **26**, 599-605.
27. Famini, G.R., Wilson, L.Y., and DeVito, S.C. (1994). Modeling cytochrome P-450

mediated acute nitrile toxicity using theoretical linear solvation energy relationships, in *Biomarkers of Human Exposures to Pesticides* (Famini *et al.*, Eds.). American Chemical Society, pp. 22-36.

28. Basak, S.C., Gute, B.D., and Grunwald, G.D. (1996). A comparative study of topological and geometrical parameters in estimating normal boiling point and octanol/water partition coefficient. *J. Chem. Inf. Comput. Sci.* **36**, 1054-1060.
29. Basak, S.C., Gute, B.D., and Grunwald, G.D. (1997). Use of topostructural, topochemical, and geometric parameters in the prediction of vapor pressure: a hierarchical QSAR approach. *J. Chem. Inf. Comput. Sci.*, in press.
30. Basak, S.C., Gute, B.D., and Grunwald, G.D. (1997). The relative effectiveness of topological, geometrical, and quantum chemical parameters in estimating mutagenicity of chemicals, in *Proceedings of the 7th International Workshop on QSARs in Environmental Sciences* (F. Chen *et al.*, Eds.) SETAC Press, in press.
31. Basak, S.C., Gute, B.D., and Grunwald, G.D. (1997). Development of a quantitative structure-activity relationship (QSAR) for estimating bioconcentration factor in *Pimephales promelas*: a hierarchical approach. In progress.
32. Hall, L.H., Kier, L.B., and Phipps, G. (1984). Structure-activity relationship studies on the toxicities of benzene derivatives: I. An additivity model. *Environ. Toxicol. Chem.* **3**, 355-365.
33. Basak, S.C., Harriss, D.K., and Magnuson, V.R. (1988). POLLY 2.3: Copyright of the University of Minnesota.
34. Wiener, H. (1947). Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **69**, 17-20.
35. Randic, M. (1975). On characterization of molecular branching. *J. Am. Chem. Soc.* **97**, 6609-6615.
36. Kier, L.B., and Hall, L.H. (1986). *Molecular Connectivity in Structure-Activity Analysis*. Research Studies Press, Letchworth, Hertfordshire, UK.
37. Basak, S.C. and Magnuson, V.R. (1988). Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math.* **19**, 17-44.
38. Raychaudhury, C., Ray, S.K., Ghosh, J.J., Roy, A.B., and Basak, S.C. (1984). Discrimination of isomeric structures using information theoretic topological indices. *J. Comput. Chem.* **5**, 581-588.
39. Bonchev, D., and Trinajstic, N. (1977). Information theory, distance matrix and molecular branching. *J. Chem. Phys.* **67**, 4517-4533.
40. Basak, S.C., Roy, A.B., and Ghosh, J.J. (1980). Study of the structure-function

relationship of pharmacological and toxicological agents using information theory, in *Proceedings of the Second International Conference on Mathematical Modelling* (X.J.R. Avula, R. Bellman, Y.L. Luke and A.K. Rigler, Eds.). University of Missouri - Rolla, pp.851-856.

41. Roy, A.B., Basak, S.C., Harriss, D.K., and Magnuson, V.R. (1984). Neighborhood complexities and symmetry of chemical graphs and their biological applications, in *Mathematical Modelling in Science and Technology* (X.J.R. Avula, R.E. Kalman, A.I. Lapis and E.Y. Rodin, Eds.). Pergamon Press, New York, pp. 745-750.
42. Balaban, A.T. (1982). Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **89**, 399-404.
43. Balaban, A.T. (1983). Topological indices based on topological distances in molecular graphs. *Pure and Appl. Chem.* **55**, 199-206.
44. Balaban, A.T. (1986). Chemical graphs. Part 48. Topological index J for heteroatom-containing molecules taking into account periodicities of element properties. *Math. Chem. (MATCH)*. **21**, 115-122.
45. Bondi, A. (1964). Van der Waals volumes and radii. *J. Phys. Chem.* **68**, 441-451.
46. Moriguchi, I., and Kanada, Y. (1977). Use of van der Waals volume in structure-activity studies. *Chem. Pharm. Bull.* **25**, 926-935.
47. Moriguchi, I., Kanada, Y., and Komatsu, K. (1976). Van der Waals volume and the related parameters for hydrophobicity in structure-activity studies. *Chem. Pharm. Bull.* **24**, 1799-1806.
48. SYBYL Version 6.1. (1994). Tripos Associates, Inc.: St. Louis, MO.
49. Mekenyany, O., Peitchev, D., Bonchev, D., Trinajstic, N., and Bangov, I. (1986). Modelling the interaction of small organic molecules with biomacromolecules. I. Interaction of substituted pyridines with anti-3-azopyridine antibody. *Arzneim.-Forsch./Drug Research* **36**, 176-183.
50. CONCORD Version 3.0.1. (1993). Tripos Associates, Inc.: St. Louis, MO.
51. Stewart, J.J.P. (1990). MOPAC Version 6.00. QCPE #455. Frank J Seiler Research Laboratory: US Air Force Academy, CO.
52. SAS Institute Inc. (1988). In *SAS/STAT User's Guide, Release 6.03 Edition*. SAS Institute Inc.: Cary, NC.
53. Topliss, J.G., and Edwards, R.P. (1979). Chance factors in studies of quantitative structure-activity relationships. *J. Med. Chem.* **22**, 1238-1244.

Table I. Sixty-nine benzene derivatives and their fathead minnow toxicities, expressed as $-\log(LC_{50})$.

No.	Compound	$-\log(LC_{50})$ (obs.)	$-\log(LC_{50})$ (est. eq. 4)	Residual
1	Benzene	3.40	3.42	-0.02
2	Bromobenzene	3.89	3.77	0.12
3	Chlorobenzene	3.77	3.75	0.02
4	Phenol	3.51	3.38	0.13
5	Toluene	3.32	3.66	-0.34
6	1,2-dichlorobenzene	4.40	4.29	0.11
7	1,3-dichlorobenzene	4.30	4.37	-0.07
8	1,4-dichlorobenzene	4.62	4.51	0.11
9	2-chlorophenol	4.02	3.79	0.23
10	3-chlorotoluene	3.84	3.88	-0.04
11	4-chlorotoluene	4.33	3.87	0.46
12	1,3-dihydroxybenzene	3.04	3.43	-0.39
13	3-hydroxyanisole	3.21	3.33	-0.12
14	2-methylphenol	3.77	3.64	0.13
15	3-methylphenol	3.29	3.60	-0.31
16	4-methylphenol	3.58	3.53	0.05
17	4-nitrophenol	3.36	3.61	-0.25
18	1,4-dimethoxybenzene	3.07	3.28	-0.21
19	1,2-dimethylbenzene	3.48	3.93	-0.45
20	1,4-dimethylbenzene	4.21	3.87	0.34
21	2-nitrotoluene	3.57	3.66	-0.09
22	3-nitrotoluene	3.63	3.53	0.10
23	4-nitrotoluene	3.76	3.49	0.27
24	1,2-dinitrobenzene	5.45	5.24	0.21
25	1,3-dinitrobenzene	4.38	4.18	0.20

26	1,4-dinitrobenzene	5.22	4.94	0.28
27	2-methyl-3-nitroaniline	3.48	3.79	-0.31
28	2-methyl-4-nitroaniline	3.24	3.51	-0.27
29	2-methyl-5-nitroaniline	3.35	3.68	-0.33
30	2-methyl-6-nitroaniline	3.80	3.84	-0.04
31	3-methyl-6-nitroaniline	3.80	3.78	0.02
32	4-methyl-2-nitroaniline	3.79	3.80	-0.01
33	4-hydroxy-3-nitroaniline	3.65	3.61	0.04
34	4-methyl-3-nitroaniline	3.77	3.73	0.04
35	1,2,3-trichlorobenzene	4.89	4.89	-0.00
36	1,2,4-trichlorobenzene	5.00	5.04	-0.04
37	1,3,5-trichlorobenzene	4.74	5.11	-0.37
38	2,4-dichlorophenol	4.30	4.33	-0.03
39	3,4-dichlorotoluene	4.74	4.26	0.48
40	2,4-dichlorotoluene	4.54	4.36	0.18
41	4-chloro-3-methylphenol	4.27	3.87	0.40
42	2,4-dimethylphenol	3.86	3.76	0.10
43	2,6-dimethylphenol	3.75	3.80	-0.05
44	3,4-dimethylphenol	3.90	3.80	0.10
45	2,4-dinitrophenol	4.04	4.14	-0.10
46	1,2,4-trimethylbenzene	4.21	4.09	0.12
47	2,3-dinitrotoluene	5.01	5.20	-0.19
48	2,4-dinitrotoluene	3.75	4.10	-0.35
49	2,5-dinitrotoluene	5.15	4.84	0.31
50	2,6-dinitrotoluene	3.99	4.41	-0.42
51	3,4-dinitrotoluene	5.08	5.11	-0.03
52	3,5-dinitrotoluene	3.91	4.05	-0.14
53	1,3,5-trinitrobenzene	5.29	5.37	-0.08
54	2-methyl-3,5-dinitroaniline	4.12	4.13	-0.01

55	2-methyl-3,6-dinitroaniline	5.34	4.80	0.54
56	3-methyl-2,4-dinitroaniline	4.26	4.28	-0.02
57	5-methyl-2,4-dinitroaniline	4.92	4.14	0.78
58	4-methyl-2,6-dinitroaniline	4.21	4.67	-0.46
59	5-methyl-2,6-dinitroaniline	4.18	4.80	-0.62
60	4-methyl-3,5-dinitroaniline	4.46	4.34	0.12
61	2,4,6-tribromophenol	4.70	4.89	-0.19
62	1,2,3,4-tetrachlorobenzene	5.43	5.62	-0.19
63	1,2,4,5-tetrachlorobenzene	5.85	5.80	0.05
64	2,4,6-trichlorophenol	4.33	4.79	-0.46
65	2-methyl-4,6-dinitrophenol	5.00	4.21	0.79
66	2,3,6-trinitrotoluene	6.37	6.36	0.01
67	2,4,6-trinitrotoluene	4.88	5.16	-0.28
68	2,3,4,5-tetrachlorophenol	5.72	5.36	0.36
69	2,3,4,5,6-pentachlorophenol	6.06	6.03	0.03

Table II. Symbols and definitions of topological and geometrical parameters.

I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
\bar{I}_D^W	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
\overline{IC}	Information content of the distance matrix partitioned by frequency of occurrences of distance h
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
O	Order of neighborhood when IC reaches its maximum value for the hydrogen-filled graph
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
IC_r	Mean information content or complexity of a graph based on the r^{th} ($r = 0-5$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r^{th} ($r = 0-5$) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r^{th} ($r = 0-5$) order neighborhood of vertices in a hydrogen-filled graph
$^h\chi$	Path connectivity index of order $h = 0-6$
$^h\chi_c$	Cluster connectivity index of order $h = 3, 5$
$^h\chi_{ch}$	Chain connectivity index of order $h = 6$
$^h\chi_{pc}$	Path-cluster connectivity index of order $h = 4-6$
$^h\chi^b$	Bond path connectivity index of order $h = 0-6$
$^h\chi_c^b$	Bond cluster connectivity index of order $h = 3, 5$
$^h\chi_{ch}^b$	Bond chain connectivity index of order $h = 6$
$^h\chi_{pc}^b$	Bond path-cluster connectivity index of order $h = 4-6$

${}^h\chi^v$	Valence path connectivity index of order $h = 0-6$
${}^h\chi_c^v$	Valence cluster connectivity index of order $h = 3, 5$
${}^h\chi_{pc}^v$	Valence path-cluster connectivity index of order $h = 4-6$
P_h	Number of paths of length $h = 1-9$
J	Balaban's J index based on distance
J^B	Balaban's J index based on bond types
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii
V_w	van der Waals volume
${}^{3D}W$	3-D Wiener number for the hydrogen-suppressed geometric distance matrix
${}^{3D}W_H$	3-D Wiener number for the hydrogen-filled geometric distance matrix

Table III. Classification of parameters used in developing models for acute aquatic toxicity (LC_{50}) in *Pimephales promelas*.

Topological	Topochemical	Geometric	Quantum Chemical AM1
I_D^W	I_{ORB}	V_w	E_{HOMO}
I_D^W	$IC_0 - IC_5$	^{3D}W	E_{HOMO1}
W	$SIC_0 - SIC_5$	$^{3D}W_H$	E_{LUMO}
I^D	$CIC_0 - CIC_5$		E_{LUMO1}
H^V	$^0\chi^b - ^6\chi^b$		ΔHf
H^D	$^3\chi_c^b$ and $^5\chi_c^b$		μ
IC	$^6\chi_{ch}^b$		
O	$^4\chi_{PC}^b - ^6\chi_{PC}^b$		
M_1	$^0\chi^v - ^6\chi^v$		
M_2	$^3\chi_c^v$ and $^5\chi_c^v$		
$^0\chi - ^6\chi$	$^4\chi_{PC}^v - ^6\chi_{PC}^v$		
$^3\chi_c$ and $^5\chi_c$	J^B		
$^6\chi_{ch}$	J^X		
$^4\chi_{PC} - ^6\chi_{PC}$	J^Y		
$P_1 - P_9$			
J			

Table IV. Summary of the regression results for all models for the full set of sixty-nine benzene derivatives.

Eq.	Parameter class	Variables Included	F	R ²	s
1	TSI	M ₁ , \overline{IC} , P ₈ , P ₉	13.3	0.453	0.58
2	TSI + TCI	M ₁ , P ₉ , SIC ₀ , CIC ₀	57.9	0.783	0.36
3	TSI + TCI + Geometric	M ₁ , P ₉ , SIC ₀ , ^{3D} W _H	61.1	0.792	0.36
4	TSI + TCI + Geometric + Quantum Chemical	M ₁ , P ₉ , SIC ₀ , ^{3D} W _H , E _{LUMO1} , ΔH_f , μ	55.0	0.863	0.30
5	TCI	SIC ₀ , SIC ₁ , CIC ₀ , ² χ^b , J^X	34.3	0.731	0.41
6	Geometric	^{3D} W _H , ^{3D} W, V _w	34.8	0.616	0.48
7	Quantum Chemical	E _{HOMO1} , E _{LUMO} , E _{LUMO1} , μ	23.8	0.598	0.50

Table V. Calculated values for the topostructural, topochemical, geometric, and quantum chemical parameters used in equation 4 (Table IV).

No.	M ₁	P _g	SIC ₀	^{3D} W _H	E _{LUMO1}	ΔH _f	μ
1	3	0	0.246	5.21	0.5540	22.0240	0.005
2	3	0	0.315	5.25	0.2447	26.7581	1.449
3	3	0	0.315	5.25	0.2632	14.8214	1.299
4	3	0	0.304	5.43	0.5095	-22.2334	1.233
5	3	0	0.227	5.79	0.5745	16.5004	0.279
6	4	0	0.341	5.28	-0.0203	9.2203	1.974
7	4	0	0.341	5.28	-0.0462	8.2544	1.218
8	4	0	0.341	5.28	-0.0988	10.4661	0.000
9	4	0	0.362	5.46	0.2406	-28.6621	0.934
10	4	0	0.284	5.81	0.2785	7.1915	1.478
11	4	0	0.284	5.82	0.3208	7.1066	1.623
12	4	0	0.323	5.64	0.3778	-66.4516	2.433
13	4	0	0.295	6.16	0.4618	-59.9961	2.338
14	4	0	0.276	5.95	0.5331	-28.9297	0.960
15	4	0	0.276	5.97	0.5610	-29.6368	1.079
16	4	0	0.276	5.97	0.4880	-29.7869	1.333
17	4	0	0.376	5.84	-0.4095	-19.5199	5.261
18	4	0	0.274	6.59	0.5766	-52.9350	2.424
19	4	0	0.213	6.22	0.6180	7.5221	0.465
20	4	0	0.213	6.28	0.6450	6.8236	0.003
21	4	0	0.341	6.11	-0.2692	19.0823	5.015
22	4	0	0.341	6.14	-0.2921	17.6145	5.443
23	4	0	0.341	6.15	-0.2334	17.2948	5.728
24	4	2	0.389	5.99	-1.2793	38.6210	7.804
25	4	0	0.389	6.01	-1.5339	33.1466	4.845
26	4	0	0.389	6.02	-1.0875	33.2941	0.013

27	4	0	0.344	6.38	-0.1596	20.4489	5.727
28	4	0	0.344	6.41	-0.0919	14.3213	7.434
29	4	0	0.344	6.41	-0.1084	19.7541	6.185
30	4	0	0.344	6.39	-0.0006	13.8471	5.374
31	4	0	0.344	6.42	0.1022	12.9086	5.649
32	4	0	0.344	6.42	0.0314	13.3128	5.280
33	4	0	0.376	6.15	-0.2384	-15.9560	6.801
34	4	0	0.344	6.41	-0.1379	18.0141	5.596
35	4	0	0.349	5.31	-0.3391	4.2313	2.070
36	4	0	0.349	5.31	-0.2761	2.9490	1.033
37	4	0	0.349	5.31	-0.3927	2.2158	0.020
38	4	0	0.385	5.49	-0.1034	-35.1296	0.395
39	4	0	0.312	5.84	0.0251	1.5862	2.296
40	4	0	0.312	5.84	0.0006	1.2199	1.464
41	4	0	0.326	5.99	0.2063	-36.1532	1.059
42	4	0	0.255	6.40	0.5006	-36.4200	1.052
43	4	0	0.255	6.38	0.5503	-35.5810	1.199
44	4	0	0.255	6.38	0.5387	-36.6403	1.229
45	4	0	0.383	6.17	-1.5210	-8.7887	6.201
46	4	0	0.202	6.64	0.6477	-0.1093	0.274
47	4	2	0.365	6.40	-1.2262	31.8226	7.909
48	4	0	0.365	6.43	-1.4332	26.3804	5.390
49	4	0	0.365	6.42	-1.0421	26.9397	0.797
50	4	0	0.365	6.39	-1.4076	30.3487	3.639
51	4	2	0.365	6.43	-1.1564	32.0703	8.256
52	4	0	0.365	6.44	-1.4923	25.3294	5.321
53	4	0	0.378	6.33	-2.5221	44.8961	0.032
54	4	0	0.362	6.66	-1.2453	27.9172	6.590
55	4	0	0.362	6.65	-0.6994	25.1359	3.166

56	4	0	0.362	6.65	-1.1532	23.8377	5.797
57	4	0	0.362	6.67	-1.3084	51.2351	7.196
58	4	0	0.362	6.68	-1.0204	18.0757	2.366
59	4	0	0.362	6.66	-1.0160	54.7718	3.199
60	4	0	0.362	6.66	-1.2172	29.5227	5.090
61	4	0	0.392	5.54	-0.4993	-2.2014	1.096
62	4	0	0.341	5.34	-0.5585	-0.5979	1.616
63	4	0	0.341	5.34	-0.6587	3.2072	0.000
64	4	0	0.392	5.52	-0.3777	-38.2930	1.083
65	4	0	0.362	6.56	-1.5102	-19.8380	4.669
66	4	2	0.365	6.66	-1.9189	46.0695	3.518
67	4	0	0.365	6.67	-2.3240	41.4239	1.418
68	4	0	0.385	5.54	-0.5526	-43.2613	1.231
69	4	0	0.362	5.57	-0.7546	-44.7215	1.238

CHARACTERIZATION OF MOLECULAR STRUCTURES
USING TOPOLOGICAL INDICES

Subhash C. Basak*

and

Brian D. Gute

Natural Resources Research Institute
University of Minnesota
5013 Miller Trunk Highway
Duluth, MN 55811

*Is all that we see or seem
But a dream within a dream?*

EDGAR ALLAN POE

SAR and QSAR in Environmental Research, in press, 1997.

*Author to whom all correspondence should be addressed.

ABSTRACT

The characterization of molecular structure using structural invariants has increased greatly over the last ten years. Specifically, topological indices have become more widely in the quantification of molecular structure for use in quantitative structure-activity relationship studies, chemical documentation, and molecular similarity studies. The basis, calculation, and utility of topological indices has been examined, with an eye to the specific advantages and problems in their use. In addition, variable clustering and principal component analysis are examined as two potential solutions to the problem of index intercorrelation.

KEYWORDS

topological indices; molecular structure; graph theory; graph invariants; variable clustering; principal component analysis

INTRODUCTION

An important area of research in computational and mathematical chemistry is the characterization of molecular structure using structural invariants.¹⁻¹⁴ The impetus for this research trend comes from various directions. Researchers in chemical documentation have searched for a set of invariants which will be more convenient than the adjacency matrix (or connection table) for the storage and comparison of chemical structures.¹⁵ Invariants have been used to order sets of molecules.^{3-5, 8, 16} With the substantial increase in available databases of chemical structures and properties, attempts have been made to develop structure-activity relationships (SARs) whereby existing molecules can be compared with other molecules (real or hypothetical) on the basis of these structural invariants. The properties of the molecules of interest can then be predicted based on molecular structure without the need for experimental data.

In this age of combinatorial chemistry thousands of molecules of known structure can be produced rapidly. However, at the same time resources for determining even the simplest properties of all of these molecules in the laboratory are unavailable. In the USA, the Toxic Substances Control Act (TSCA) Inventory includes nearly 74,000 chemicals and the list is growing at a rate of more than 2,000 new submissions to the United States Environmental Protection Agency (USEPA) for the Premanufacture Notification (PMN) process per year.¹⁷⁻²⁰ At present, risk assessment of the PMN chemicals is carried out using limited

test data. For example, approximately 15% of PMN submissions have empirical mutagenicity data. Under such circumstances, structural descriptors will play a pivotal role in comparing molecules with one another and in predicting their properties.

MOLECULAR STRUCTURE – Beauty in the Eye of the Beholder or Conundrum?

The main hurdle to the characterization of molecular structure is the lack of uniformity in its definition and quantification. The term *molecular structure* represents a set of nonequivalent and probably disjoint concepts.²¹ For example, the term “molecule” means different things when it represents an assembly of identifiable atoms held together by fairly rigid bonds as compared to a collection of delocalized nuclei and electrons in which all identical particles are indistinguishable.²¹ There is no reason to believe that when we discuss diverse topics (e.g., chemical synthesis, reaction rates, spectroscopic transitions, reaction mechanisms, and *ab initio* calculations) using the notion of *molecular structure*, that the different meanings we attach to this term originate from the same fundamental concept.^{21,22} This fundamental problem has been described succinctly by Woolley.²²

“...there is no reason to suppose that the same basic idea can provide a basis for the discussion of all molecular experiments.

This is understandable if one recognizes that every physical and chemical concept is only defined with respect to a certain class of experiments, so that it is perfectly reasonable for different sets of concepts, although mutually incompatible, to be applicable to different experiments."

In the context of molecular science, the various concepts of molecular structure (e.g., classical valence bond representation, various chemical graph-theoretic representations, the ball-and-stick model, representation by minimum energy conformation, semi-symbolic contour maps, or symbolic representation by Hamiltonian operators) are distinct molecular models derived through different means of abstraction from the same chemical reality or molecule.²³ In each instance, the equivalence class (concept or model of molecular structure) is generated by selecting certain aspects while ignoring other unique properties of those actual events. This explains the plurality of the concepts of molecular structure and their autonomous nature, the word autonomous being used in the sense that one concept is not logically derived from the other.

GRAPHS AND MOLECULAR STRUCTURE

At the most fundamental level, the structural model of an assembled entity (e.g., a molecule consisting of atoms) may be defined as the pattern of relationship

among its parts as distinct from the values associated with them.²⁴ Constitutional formulae of molecules are graphs where vertices represent the set of atoms and edges represent chemical bonds.²⁵ The pattern of connectedness of atoms in a molecule is preserved by constitutional graphs. A graph (more correctly a non-directed graph) $G = [V, E]$ consists of a finite nonempty set V of points together with a prescribed set E of unordered pairs of distinct points of V .²⁶ A *structural model* assigns to the points of G a realization in some applied field and each element of E indicates a pair of entities (elements of the structural model) which are in the finite nonempty irreflexive symmetric binary relation described by G . For example, when elements of the set V symbolize atomic cores without valence electrons and the elements of E represent covalent two-electron bonds, G is the molecular graph or constitutional graph of a covalent chemical species. Such a graph can represent structural formulae of a large number of organic compounds. Since more than 90% of chemical compounds described so far are either organic or contain organic ligands, such a graph has been found to be useful in chemistry.¹³ The edge set need not always represent a covalent bond. In fact, elements of E may symbolize almost any type of bond (e.g., ionic, coordinate, hydrogen, or weak bonds representing transition states of an SN_2 reaction, etc.).²⁷⁻²⁹ If the interaction between a pair of atoms is asymmetric (e.g., in case of sufficiently polar covalent bonds, hydrogen bond donor acidity, hydrogen bond acceptor basicity, or charge transfer complex formation) the bonding pattern can be represented by a binary relation which is anti-reflexive

and asymmetric.⁶ Further refinement could be achieved through the assignment of weights to the vertices or edges,³ and use of multiple edges between a pair of atoms held together both by *sigma* and *pi* bonds. The weighted pseudograph appears to be the most general model capable of symbolizing the bonding pattern of a large number of organic and inorganic chemicals.

For a long time, chemists have relied on visual perception to relate various aspects of constitutional graphs to observable phenomena. The power of graph-theoretic formalism in chemistry is evident from its successful applications in chemical documentation, isomer discrimination and characterization of molecular branching, enumeration of constitutional isomers associated with a particular empirical formula, calculation of quantum chemical parameters, structure-physicochemical property correlations, and chemical structure-biological activity relationships.³⁰⁻³⁷

GRAPHS AS MOLECULAR MODELS

Any concept of molecular structure is a hypothetical sketch of the organization of atoms within the molecule. Such a *model object* is a general theory and remains empirically untestable. A model object has to be grafted to a specific theory to generate a *theoretical model* which can be empirically tested.³⁸ For example, when it was suggested by Sylvester in 1878 that the structural formula of a molecule is a special kind of graph, it was an innovative general theory without

any predictive potential.³⁹ When the idea of combinatorics was applied on chemical graphs (model object), it could be predicted that "there should be exactly two isomers of butane (C_4H_{10})" because "there are exactly two tree graphs with four vertices" when one considers only the non-hydrogen atoms present in C_4H_{10} .¹³ This is a theoretical model of limited predictive potential. Although it predicts the existence of chemical species, given a set of molecules (e.g., isomers of hexane [C_6H_{14}]) the model is incapable of predicting any properties for these molecules. This is due to the fact that any empirical property P maps a set of chemical structures into the set R of real numbers and thereby orders the set empirically. Therefore, to predict the property from structure, we need a nonempirical (structural) ordering scheme which closely resembles the empirical ordering of structures as determined by P . This is a more specific theoretical model based on the same model object (chemical graph) and can be accomplished by using specific graph invariant(s).

CHARACTERIZATION OF MOLECULAR GRAPHS

Molecular graphs can be characterized by graph invariants. A graph invariant is a graph-theoretic property which is preserved by isomorphism.²⁶ A graph invariant could be a polynomial, a sequence of numbers, or a single number. The characteristic polynomial of a graph and the spectra of graphs are graph invariants. Numerical graph invariants derived from molecular graphs are called

graph-theoretic indices or topological indices.²⁵ Topological indices quantitatively describe molecular topology and are sensitive to such structural attributes as size, shape, patterns of branching, bonding types, and cyclicity of molecules.

Topological indices (TIs) can sometimes be derived conveniently from different matrices such as the adjacency matrix and the distance matrix. The origins of such TIs illuminate the fundamental structural features that they quantify. On the other hand, some indices are derived to quantify a key structural feature which is qualitative and only understood intuitively. In deriving his original connectivity index ($'x$), Randić asked the question: which of the two heptane isomers, *viz.*, 3-methylhexane and 3-ethyl pentane, is more branched.⁹ Until that time, branching was understood only intuitively; Randić derived a quantitative description of branching based on the graph-theoretic treatment of the structures. In addition, information theoretic indices of chemical structures have been derived to answer the question: which of a collection of structures is more complex or heterogeneous? Different measures of molecular complexity attempt to answer this question from different points of view.⁴⁰ In the following section we discuss the structural basis and method of calculation for some of the major topological indices.

CALCULATION OF TOPOLOGICAL INDICES

The Wiener index (W),⁴¹ the first topological index reported in the chemical literature, may be calculated from the distance matrix $D(G)$ of a hydrogen-suppressed chemical graph G as the sum of the entries in the upper triangular distance submatrix. The distance matrix $D(G)$ of a nondirected graph G with n vertices is a symmetric $n \times n$ matrix (d_{ij}), where d_{ij} is equal to the distance between vertices v_i and v_j in G . Each diagonal element d_{ii} of $D(G)$ is zero. We give below the distance matrix $D(G_1)$ of the labelled hydrogen-suppressed graph G_1 of 2,3-dimethylhexane (Fig. 1):

$$D(G_1) = \begin{matrix} & (1) & (2) & (3) & (4) & (5) & (6) & (7) & (8) \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{matrix} & \left[\begin{matrix} 0 & 1 & 2 & 2 & 3 & 3 & 4 & 5 \\ 1 & 0 & 1 & 1 & 2 & 2 & 3 & 4 \\ 2 & 1 & 0 & 2 & 3 & 3 & 4 & 5 \\ 2 & 1 & 2 & 0 & 1 & 1 & 2 & 3 \\ 3 & 2 & 3 & 1 & 0 & 2 & 3 & 4 \\ 3 & 2 & 3 & 1 & 2 & 0 & 1 & 2 \\ 4 & 3 & 4 & 2 & 3 & 1 & 0 & 1 \\ 5 & 4 & 5 & 3 & 4 & 2 & 1 & 0 \end{matrix} \right] \end{matrix}$$

W is calculated as:

$$W = \frac{1}{2} \sum_{ij} d_{ij} = \sum_h h \cdot g_h \quad (1)$$

where g_h is the number of unordered pairs of vertices whose distance is h . Thus for $D(G_1)$, W has a value of seventy.

[Insert Fig. 1 here]

Randić's connectivity index,⁹ and higher-order connectivity path, cluster, path-cluster and chain types of simple, bond and valence connectivity parameters were calculated using the method of Kier and Hall.¹⁰ P_h parameters, number of paths of length h ($h = 0, 1, \dots, 10$) in the hydrogen-suppressed graph, are calculated using standard algorithms.

Balaban defined a series of indices based upon distance sums within the distance matrix for a chemical graph which he designated as J indices.⁴²⁻⁴⁴ These indices are highly discriminating with low degeneracy. Unlike W , the J indices range of values are independent of molecular size.

Information-theoretic topological indices are calculated by the application of information theory on chemical graphs. An appropriate set A of n elements is derived from a molecular graph G depending upon certain structural characteristics. On the basis of an equivalence relation defined on A , the set A is partitioned into disjoint subsets A_i of order n_i ($i = 1, 2, \dots, h$; $\sum_i n_i = n$). A probability distribution is then assigned to the set of equivalence classes:

A_1, A_2, \dots, A_h

p_1, p_2, \dots, p_h

where $p_i = n_i / n$ is the probability that a randomly selected element of A will occur in the i^{th} subset.

The mean information content of an element of A is defined by Shannon's relation:⁴⁵

$$IC = - \sum_{i=1}^n p_i \log_2 p_i \quad (2)$$

The logarithm is taken at base 2 for measuring the information content in bits.

The total information content of the set A is then $n \times IC$.

It is to be noted that the information content of a graph G is not uniquely defined. It depends on how the set A is derived from G as well as on the equivalence relation which partitions A into disjoint subsets A_i . For example, when A constitutes the vertex set of a chemical graph G , two methods of partitioning have been widely used: a) chromatic-number coloring of G where two vertices of the same color are considered equivalent, and b) determination of the orbits of the automorphism group of G thereafter vertices belonging to the same orbit are considered equivalent.

Rashevsky was the first to calculate the information content of graphs where "topologically equivalent" vertices were placed in the same equivalence class.⁴⁶ In Rashevsky's approach, two vertices u and v of a graph are said to be topologically equivalent if and only if for each neighboring vertex u_i ($i = 1, 2, \dots, k$) of the vertex u , there is a distinct neighboring vertex v_i of the same degree for the vertex v . While Rashevsky used simple linear graphs with indistinguishable

vertices to symbolize molecular structure, weighted linear graphs or multigraphs are better models for conjugated or aromatic molecules because they more properly reflect the actual bonding patterns, *i.e.*, electron distribution.

To account for the chemical nature of vertices as well as their bonding pattern, Sarkar *et al.* calculated information content of chemical graphs on the basis of an equivalence relation where two atoms of the same element are considered equivalent if they possess an identical first-order topological neighborhood.⁴⁷ Since properties of atoms or reaction centers are often modulated by stereo-electronic characteristics of distant neighbors, *i.e.*, neighbors of neighbors, it was deemed essential to extend this approach to account for higher-order neighbors of vertices. This can be accomplished by defining open spheres for all vertices of a chemical graph. If r is any non-negative real number and v is a vertex of the graph G , then the open sphere $S(v, r)$ is defined as the set consisting of all vertices v' in G such that $d(v, v') < r$. Therefore, $S(v, 0) = \phi$, $S(v, r) = v$ for $0 < r < 1$, and $S(v, r)$ is the set consisting of v and all vertices v' of G situated at unit distance from v , if $1 < r < 2$.

One can construct such open spheres for higher integral values of r . For a particular value of r , the collection of all such open spheres $S(v, r)$, where v runs over the whole vertex set V , forms a neighborhood system of the vertices of G . A suitably defined equivalence relation can then partition V into disjoint subsets consisting of vertices which are topologically equivalent for r^{th} order neighborhood. Such an approach has been developed and the information-

theoretic indices calculated based on this idea are called indices of neighborhood symmetry.⁴⁰

In this method, chemicals are symbolized by weighted linear graphs. Two vertices u_o and v_o of a molecular graph are said to be equivalent with respect to r^{th} order neighborhood if and only if corresponding to each path u_o, u_1, \dots, u_r of length r , there is a distinct path v_o, v_1, \dots, v_r of the same length such that the paths have similar edge weights, and both u_o and v_o are connected to the same number and type of atoms up to the r^{th} order bonded neighbors. The detailed equivalence relation has been described in earlier studies.^{40, 48}

Once partitioning of the vertex set for a particular order of neighborhood is completed, IC_r is calculated by Eq. 2. Subsequently, Basak *et al.* defined another information-theoretic measure, structural information content (SIC_r), which is calculated as:

$$SIC_r = IC_r / \log_2 n \quad (3)$$

where IC_r is calculated from Eq. 2 and n is the total number of vertices of the graph.⁴⁹

Another information-theoretic invariant, complementary information content (CIC_r), is defined as:

$$CIC_r = \log_2 n - IC_r \quad (4)$$

C/C_r represents the difference between maximum possible complexity of a graph (where each vertex belongs to a separate equivalence class) and the realized topological information of a chemical species as defined by IC_r .⁵⁰

In Fig. 2, the calculation of IC_r , S/C , and C/C , is demonstrated for the hydrogen-filled graph (G_h) of 2,3-dimethylhexane.

[Insert Fig. 2 here]

The information-theoretic index on graph distance, I_D^W is calculated from the distance matrix $D(G)$ of a chemical graph G as follows:¹¹

$$I_D^W = W \log_2 W - \sum_h g_h \cdot h \log_2 h \quad (5)$$

The mean information index, \bar{I}_D^W , is found by dividing the information index I_D^W by W . The information theoretic parameters defined on the distance matrix, H^P and H^V , were calculated by the method of Raychaudhury *et al.*¹²

THEORETICAL METHODS

Databases and Calculations

Two data sets were used for this study: the first consists of the seventy-four

alkanes (C_2 - C_9) and the second, more heterogeneous set was taken from the STARLIST group of chemicals.⁵¹ The STARLIST subset includes 219 chemicals for which HB_1 was equal to zero and calculated $\log P$ values fell in the range of -2 to 5.5. HB_1 is a measure of the hydrogen bonding potential of a chemical. Chemical structures for these compounds were encoded using the SMILES line notation for chemical structures and entered into the computer program POLLY version 2.3 for the calculation of indices.⁵² Table I provides a comprehensive list and brief descriptions for these indices.

Statistical Methods

Initially all TIs were transformed by the natural logarithm of the index plus one. This is routinely done to scale the indices since there may be a difference of several orders of magnitude between indices and some may equal zero.

From the original sets of 102 indices calculated for both data sets, it was necessary to remove some indices. Some of the indices for the set of alkanes (e.g., the simple, valence, and bond connectivity indices) were completely redundant. Other indices were removed because they had values of zero for all compounds. This "cleaning" of the sets of TIs left fifty-three indices for the alkanes and ninety-eight indices for the STARLIST set.

Variable clustering and principal component analysis were used on the remaining indices to minimize problems of intercorrelation amongst the indices.

The variable clustering was conducted using the SAS procedure VARCLUS which divides the indices into disjoint clusters which are essentially unidimensional based on the correlation matrix.⁵³ From each cluster, the index which was most correlated with the cluster was selected as the best representative of that cluster. In this way, individual indices are retained while minimizing intercorrelations. This procedure resulted in the retention of eight TIs for the alkanes; H^V , SIC_0 , SIC_1 , SIC_4 , 3X_c , 5X_c , P_4 , P_6 ; and twelve TIs for the STARLIST data; I^D_w , IC_4 , SIC_3 , CIC_1 , 4X , $^4X_{ch}$, $^6X^v_{ch}$, $^3X^b_c$, $^5X^b_c$, $^3X^b_{pc}$, P_8 , J^B . TI values for a subset of the alkanes, the eighteen octane isomers, are presented in Table II.

The principal component analysis (PCA) was accomplished using the SAS procedure PRINCOMP. The PCA produces linear combinations of the TIs, called principal components (PCs) which are derived from the correlation matrix.⁵⁴ The first PC has the largest variance, or eigenvalue, of the linear combination of TIs. Each subsequent PC explains the maximal index variance orthogonal to previous PCs, eliminating the redundancy which can occur with TIs. The maximum number of PCs generated is equal to the number of individual TIs available. For the purposes of this study, only PCs with eigenvalues greater than one were retained. A more detailed explanation of this approach has been provided in a previous study by Basak *et al.*³ The seven PCs with eigenvalues greater than one and the ten PCs with eigenvalues greater than one were retained for the alkanes and STARLIST set respectively. Table III presents the

PCs for the octane isomers, a subset of the seventy-four alkanes.

DISCRIMINATION OF ISOMERS USING TOPOLOGICAL INDICES AND PRINCIPAL COMPONENTS DERIVED FROM THEM

Topological aspects of chemicals have been used in chemical documentation.

One line of research in this area has been the development of topological indices which are more discriminatory. For example, the J index developed by Balaban is one of the most discriminatory indices. Randić developed the concept of molecular identification number (I.D. number) by combining a few topological aspects of structures. Other authors have used more than one index for this purpose. One example is the topological superindex proposed by Bonchev *et al.* where they use a collection of indices as the superindex.⁵⁵ Two structures are said to be distinct if the magnitudes of any one of the component indices differ for them.

In view of the intercorrelation of indices and the fact that a large number of TIs have been defined in the literature, we have been interested in deriving orthogonal parameters from TIs. We have employed two statistical methods: variable clustering and principal components analysis (PCA). In the former method, we begin with the TIs calculated by POLLY and derive a small set of original variables which are minimally intercorrelated. In the case of seventy-four alkanes the method retained eight indices. In case of PCA, seven principal

components (PCs) are derived from original variables and these PCs are linear combinations of all the TIs. For the STARLIST set, twelve TIs were retained by variable clustering, while ten PCs were derived.

We were interested to see the discriminatory power of the TIs selected by variable clustering *vis-a-vis* the PCs. Values of the TIs selected by the variable clustering technique and the first seven PCs with eigenvalue greater than 1.0 for the set of eighteen octane isomers are presented in Tables II and III respectively. It is clear from the data that some individual TIs are not sufficiently discriminatory for the eighteen octane isomers. On the other hand, each PC is unique for any given structure, making them more discriminatory than any individual TI. In the interest of space, the values of the TIs and PCs for all of the alkanes and for the STARLIST set were not included in the tables, however, this information is available upon request from the authors.

TOPOLOGICAL INDEX SPACE VIS-A-VIS PC SPACE: What Do They Mean?

Each TI quantifies certain aspects of molecular structure. Distinct indices selected by the variable clustering procedure encode different information regarding molecular structure (model object). For example, indices like the connectivity index or Wiener index quantify adjacency information of the simple planar graph model of molecules. On the other hand, information theoretic graph invariants quantify the degree of complexity of the molecular graph. Intuitively,

these are distinct aspects of molecular structure and this notion is borne out by the result of variable clustering analysis on the set of 102 TIs calculated by POLLY. It is tempting to speculate that each index retained by variable clustering represents one distinct aspect of molecular architecture and that, collectively, the TIs form the structure space of the set of chemicals. Such a space can be used for the discrimination of structures and structure-property correlation. The magnitudes of eight TIs for the eighteen octane isomers show that the TIs selected by variable clustering have reasonable power for discriminating isomeric structures.

At the level of PCs, we have derived a certain number of orthogonal variables using PCA of the indices. For the alkanes we had seven PCs with eigenvalues greater than 1.0 (Table III) whereas for the structurally diverse set of 219 compounds we had ten PCs with eigenvalues greater than 1.0. This result indicates that the structure space for the set of 219 molecules is more complex than that for the set of seventy-four alkanes. This is in agreement with our intuitive notion that molecules with heteroatoms and many functional groups are more complex than molecules devoid of any heteroatom. Finally, the pattern of correlation of the individual PCs with the TIs can help us in understanding the nature of the axes derived by PCA (Tables IV and V).

DISCUSSION

The major objectives of this paper were: a) to illuminate the fundamental nature of mathematical invariants of molecular structure, b) to study the utility of graph invariants in the characterization of molecular structure, and c) to study the intercorrelation of indices and extraction of orthogonal variables from TIs.

It is clear from the results presented in this paper that the various classes of mathematical invariants quantify different aspects of molecular architecture. They depend principally on the structural model (model object) used for the calculation of the invariant as well as the intuitive aspect of molecular structure they are used to quantify. For example, connectivity indices and neighbor complexity indices were designed to quantify distinct aspects of molecular structure. The results of variable clustering of the congeneric set of alkanes and the diverse set of 219 chemicals show that these indices encode largely independent structural information about these molecules.

Many structural schemes have been developed for the derivation of numbers or sets of numbers which can discriminate closely related structures so that they can be useful in chemical documentation. The results presented in this paper show that both the collection of indices selected by variable clustering as well as the PCs can discriminate among the eighteen octane isomers (Tables II-V). It is also clear from the data that the PCs are more discriminatory than the individual indices. For example, each PC has distinct values for all eighteen octane isomers. PCs derived from TIs have also been used in the discrimination of isospectral molecular graphs where individual indices show a high degree of

degeneracy.⁵⁶

Variable clustering of TIs for the set of seventy-four alkanes retained eight parameters which can be classified into three subsets: a) H^V , P_4 , and P_8 which represent generalized size and shape; b) SIC_0 , SIC_1 , and SIC_4 which quantify molecular complexity; and c) 3X_c and 5X_c which encode information about molecular branching. In the case of the more diverse set of 219 chemicals, the indices retained after variable clustering fall into four subclasses: a) I^D_w , P_8 , and 4X (general shape and size); b) IC_4 , SIC_3 , and CIC_1 (complexity); c) $^4X_{ch}$ and $^6X^V_{ch}$ (cyclicity); and d) $^3X_c^b$, $^5X_c^b$, $^3X_{pc}^b$, and J^B (branching). A perusal of results from both the sets indicate that distinct indices quantify different intuitive aspects of molecular structure.

A similar picture emerges from the principal component analysis of both sets of molecules. The first PC is strongly correlated with variables which quantify shape and size. The next important factor is molecular complexity which is encoded by the second PC (Tables IV and V). The higher order PCs (3-5) are strongly correlated with invariants which quantify such subtle structural factors as branching, cyclicity, etc. It may be mentioned that such a result emerged from our earlier studies on a large, diverse set of 3,692 chemicals.^{3,57}

In conclusion, mathematical invariants derived from chemical topology quantify different aspects of molecular architecture which are intuitively understood by the chemist. One can create a structure space from these invariants taking uncorrelated structural information (indices or PCs). Such

orthogonal factors can be useful in the discrimination of closely related structures like isomers and in the creation of structure spaces. Metrics defined on such spaces have been useful in the quantification of molecular similarity.^{3-5, 58-63} Orthogonal variables derived by PCA or variable clustering can also be used in QSAR studies pertaining to pharmacology and toxicology.^{1, 2, 6, 33-36, 40, 48-50, 64-68}

ACKNOWLEDGMENTS

This is contribution number XXX from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported in part by grants F49620-94-1-0401 and F49620-96-1-0330 from the United States Air Force, a grant from Exxon Corporation and the Structure-Activity Relationship Consortium (SARCON) of the Natural Resources Research Institute of the University of Minnesota. The authors would like to extend their thanks to Greg Grunwald for technical support.

REFERENCES

1. Basak, S.C., Grunwald, G.D., and Niemi, G.J. (1997). Use of graph-theoretic and geometrical molecular descriptors in structure-activity relationships, in *From Chemical Topology to Three-Dimensional to Three-Dimensional Geometry* (A. T. Balaban, Ed.). Plenum Press, New York, pp. 73-116.
2. Basak, S.C., Niemi, G. J. and Veith, G.D. (1990). Optimal characterization of structure for prediction of properties. *J. Math. Chem.* 4, 185-205.
3. Basak, S.C., Magnuson, V. R., Niemi, G. J., and Regal, R.R. (1988). Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math.* 19, 17-44.
4. Basak, S.C., Bertelsen, S., and Grunwald, G.D. (1994). Application of graph theoretical parameters in quantifying molecular similarity and structure-activity relationships. *J. Chem. Inf. Comput. Sci.* 34, 270-276.
5. Basak, S.C. and Grunwald, G.D. (1994). Use of topological space and property space in selecting structural analogs. *Mathl. Modelling and Sci. Comput.*, in press.
6. Basak, S.C., Niemi, G.J., and Veith, G.D. (1990). Recent developments in the characterization of chemical structure using graph-theoretic indices, in *Computational Chemical Graph Theory and Combinatorics* (D.H. Rouvray, Ed.). Nova, New York, pp. 235-277.
7. Fisanick, W., Cross, K.P., and Rusinko, III, A. (1992). Similarity searching on CAS registry substances. 1. Molecular property and generic atom triangle geometric searching. *J. Chem. Inf. Comput. Sci.* 32, 664-674.
8. Carhart, R.E., Smith, D.H., and Venkataraghavan, R. (1985). Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* 25, 64-73.
9. Randić, M. (1975). On characterization of molecular branching. *J. Am. Chem. Soc.* 97, 6609-6615.
10. Kier, L. B. and Hall, L.H. (1986). *Molecular Connectivity in Structure-Activity Analysis*. Research Studies Press, Letchworth, Hertfordshire, U.K.
11. Bonchev, D. and Trinajstić, N. (1977). Information theory, distance matrix and molecular branching. *J. Chem. Phys.* 67, 4517-4533.

12. Raychaudhury, C., Ray, S.K., Ghosh, J.J., Roy, A.B., and Basak, S.C. (1984). Discrimination of isomeric structures using information-theoretic topological indices. *J. Comput. Chem.* **5**, 581-588.
13. Balaban, A.T. (1985). Applications of graph theory in chemistry. *J. Chem. Inf. Comput. Sci.* **25**, 334-343.
14. Basak, S.C. and Grunwald, G.D. (1993). Use of graph invariants, volume and total surface area in predicting boiling point of alkanes. *Mathl. Modelling and Sci. Comput. Modelling* **2**, 735-740.
15. Randić, M. (1984). On molecular identification numbers. *J. Chem. Inf. Comput. Sci.* **24**, 164-175.
16. Wilkins, C.L. and Randić, M. (1980). A graph theoretical approach to structure-property and structure-activity correlations. *Theoretica Chimica Acta* **58**, 45-68.
17. Auer, C.M., Nabholz, J.V., and Baetcke, K.P. (1990). Mode of action and the assessment of chemical hazards in the presence of limited data: use of structure-activity relationships (SAR) under TSCA, Section 5. *Environ. Health Perspect.* **87**, 183-197.
18. National Research Council (NRC). (1984). *Toxicity Testing Strategies to Determine Needs and Priorities*. National Academy Press, Washington, D. C.
19. Arcos, J.C. (1987). Structure-activity relationships: criteria for predicting carcinogenic activity of chemical compounds. *Environ. Sci. Technol.* **21**, 743-745.
20. Toxic Substances Control Act (TSCA). (1976). Public Law 94-469, 90 Stat. 2003, October 11, 1976.
21. Weininger, S.J. (1984). The molecular structure conundrum: Can classical chemistry be reduced to quantum chemistry? *J. Chem. Educ.* **61**, 939-944.
22. Woolley, R.G. (1978). Must a molecule have a shape. *J. Am. Chem. Soc.* **100**, 1073-1078.
23. Primas, H. (1981). *Chemistry, Quantum Mechanics and Reductionism*. Springer-Verlag, Berlin.
24. Whyte, L.L. (1965). Atomism, structure and form: a report on the natural philosophy of form, in *Structure in Art and Science* (G. Kepes, Ed.). George

Braziler, Inc., New York, pp. 20-28.

25. Trinajstić, N. (1983). *Chemical Graph Theory*, Vols. I & II. CRC Press, Boca Raton, Florida.
26. Harary, F. (1969). *Graph Theory*. Addison Wesley Publishing Co., Reading, Massachusetts.
27. Spialter, L. (1963). The atom connectivity matrix (ACM) and its characteristic polynomial (ACMCP): a new computer-oriented chemical nomenclature. *J. Am. Chem. Soc.* 85, 2012-2013.
28. Spialter, L. (1964). The atom connectivity matrix (ACM) and its characteristic polynomial (ACMCP). *J. Chem. Doc.* 4, 261-269.
29. Spialter, L. (1964). The atom connectivity matrix characteristic polynomial (ACMCP) and its physico-geometric (topological) significance. *J. Chem. Doc.* 4, 269-274.
30. Kennedy, J.W. and Quintas, L.V. (1988). *Applications of Graphs in Chemistry and Physics*. North-Holland, Amsterdam.
31. Randić, M. (1984). Nonempirical approach to structure-activity studies. *Int. J. Quantum Chem. Quantum Biol. Symp.* 11, 137-153.
32. Sabljić, A. and Trinajstić, N. (1981). Quantitative structure-activity relationships: the role of topological indices. *Acta Pharm. Jugosl.* 31, 189-214.
33. Basak, S.C., Gieschen, D.P., Hariss, D.K., and Magnuson, V.R. (1983). Physicochemical and topological correlates of the enzymatic acetyltransfer reaction. *J. Pharm. Sci.* 72, 934-937.
34. Basak, S.C., Monsrud, L.J., Rosen, M.E., Frane, C.M., and Magnuson, V.R. (1986). A comparative study of lipophilicity and topological indices in biological correlation. *Acta Pharm. Jugosl.* 36, 81-95.
35. Basak, S.C. (1987). Use of molecular complexity indices in predictive pharmacology and toxicology: a QSAR approach. *Med. Sci. Res.* 15, 605-609.
36. Basak, S.C. (1988). Binding of barbiturates to cytochrome P₄₅₀: a QSAR study using log P and topological indices. *Med. Sci. Res.* 16, 281-282.
37. Trinajstić, N., Randić, M., and Klein, D.J. (1986). On the quantitative structure-activity relationship in drug research. *Acta Pharm. Jugosl.* 36, 267-279.

38. Bunge, M. (1973). *Method, Model and Matter*. D. Reidel Publishing Co., Dordrecht-Holland/Boston.

39. Sylvester, J.J. (1878). *Amer. J. Math.* **1**, 64.

40. Roy, A.B., Basak, S.C., Harriss, D.K., and Magnuson, V.R. (1984). Neighborhood complexities and symmetry of chemical graphs and their biological applications, in *Mathematical Modelling in Science and Technology* (X.J.R. Avula, R.E. Kalman, A.I. Liapis and E.Y. Rodin, Eds.). Pergamon Press, Elmsford, New York, pp. 745-750.

41. Wiener, H. (1947). Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **69**, 17-20.

42. Balaban, A.T. (1982). Highly discriminating distance-based topological indicex. *Chem. Phys. Lett.* **89**, 399-404.

43. Balaban, A.T. (1983). Topological indices based on topological distances in molecular graphs. *Pure and Appl. Chem.* **55**, 199-206.

44. Balaban, A.T. (1985). Chemical graphs. Part 48. Topological index J for heteroatom-containing molecules taking into account periodicities of element properties. *Math. Chem. (MATCH)* **21**, 115-122.

45. Shannon, C.E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379-423.

46. Rashevsky, N. (1955). Life, information theory and topology. *Bull. Math. Biophys.* **17**, 229-235.

47. Sarkar, R., Roy, A.B., and Sarkar, R.K. (1978). Topological information content of genetic molecules - I. *Math. Biosci.* **39**, 299-312.

48. Magnuson, V.R., Harriss, D.K., and Basak, S.C. (1983). Topological indices based on neighborhood symmetry: chemical and biological applications, in *Studies in Physical and Theoretical Chemistry* (R.B. King, Ed.). Elsevier, Amsterdam, pp. 178-191.

49. Basak, S.C., Roy, A.B., and Ghosh, J.J. (1980). Study of the structure-function relationship of pharmacological and toxicological agents using information theory, in *Proceedings of the Second International Conference on Mathematical Modelling* (X.J.R. Avula, R. Bellman, Y.L. Luke and A.K. Rigler, Eds.). University of Missouri - Rolla, pp. 851-856.

50. Basak, S.C. and Magnuson, V.R. (1983). Molecular topology and narcosis. *Arzneim-Forsch. Drug Research* 33, 501-503.
51. Leo, A. and Weininger D. (1984). *CLOGP Version 3.2 User Reference Manual*. Medicinal Chemistry Project, Pomona College, Claremont, CA.
52. Basak, S.C., Harriss, D.K., and Magnuson, V.R. (1988). POLLY 2.3: Copyright of the University of Minnesota.
53. SAS Institute Inc. (1988). In *SAS/STAT User's Guide, Release 6.03 Edition*. SAS Institute Inc., Cary, NC, Chapter 34, pp. 949-965.
54. SAS Institute Inc. (1988). In *SAS/STAT User's Guide, Release 6.03 Edition*. SAS Institute Inc., Cary, NC, Chapter 34, pp. 751-771.
55. Bonchev, D., Mekenyan, O., and Trinajstić, N. (1981). Isomer discrimination by topological information approach. *J. Comput. Chem.* 2, 127-148.
56. Balasubramanian, K. and Basak, S.C. (1997). Characterization of isospectral graphs using graph invariants and derived orthogonal parameters. *J. Chem. Inf. Comput. Sci.*, in preparation.
57. Basak, S.C., Magnuson, V.R., Niemi, G.J., Regal, R.R., and Veith, G.D. (1987). Topological indices: their nature, mutual relatedness, and applications, in *Mathematical Modelling in Science and Technology* (X.J.R. Avula, G. Leitmann, C.D. Mote, Jr. and E.Y. Rodin, Eds.). Pergamon Press: Oxford, pp. 300-305.
58. Basak, S.C., Gute, B.D., and Grunwald, G.D. (1996). Estimation of normal boiling points of haloalkanes using molecular similarity. *Croat. Chem. Acta* 69, 1159-1173.
59. Basak, S.C., Gute, B.D., and Grunwald, G.D. (1997). Development and application of molecular similarity methods: using nonempirical parameters. *Mathl. Model and Sci. Comput.*, in press.
60. Basak, S.C., and Gute, B.D. (1997). Use of graph-theoretic parameters in predicting inhibition of microsomal p-hydroxylation of aniline by alcohols: a molecular similarity approach, in *Proceedings of the 2nd International Congress on Hazardous Waste: Impact on Human and Ecological Health* (B.L. Johnson, C. Xintaras, and J.S. Andrews, Jr., Eds.). Princeton Scientific Publishing Co., Inc., New Jersey, pp. 492-504.
61. Basak, S.C., and Grunwald, G.D. (1994). Molecular similarity and risk assessment: analog selection and property estimation using graph invariants.

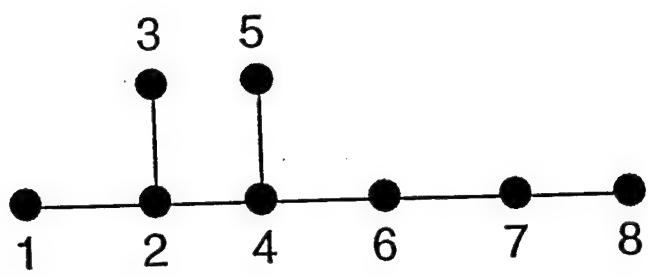
SAR QSAR Environ. Res. **2**, 289-307.

62. Basak, S.C., and Grunwald, G.D. (1995). Molecular similarity and estimation of molecular properties. *J. Chem. Inf. Comput. Sci.* **35**, 366-372.
63. Basak, S.C., and Grunwald, G.D. (1995). Tolerance space and molecular similarity. *SAR QSAR Environ. Res.* **3**, 265-277.
64. Basak, S.C., Gute, B.D., and Ghatak, S. (1997). Prediction of complement-inhibitory activity of benzamidines using topological and geometric parameters. *J. Chem. Inf. Comput. Sci.*, submitted.
65. Basak, S.C., Gute, B.D., and Grunwald, G.D. (1996). A comparative study of topological and geometrical parameters in estimating normal boiling point and octanol/water partition coefficient. *J. Chem. Inf. Comput. Sci.* **36**, 1054-1060.
66. Basak, S.C., Gute, B.D., and Grunwald, G.D. (1997). Use of topostructural, topochemical, and geometric parameters in the prediction of vapor pressure: a hierarchical QSAR approach. *J. Chem. Inf. Comput. Sci.*, submitted.
67. Basak, S.C., Gute, B.D., and Grunwald, G.D. (1997). The relative effectiveness of topological, geometrical, and quantum chemical parameters in estimating mutagenicity of chemicals, in *Proceedings of the 7th International Workshop on QSARs in Environmental Sciences* (F. Chen et al., Eds.). SETAC Press, in press.
68. Basak, S.C., Gute, B.D., and Grunwald, G.D. (1997). Development of a quantitative structure-activity relationship (QSAR) for estimating bioconcentration factor in *Pimephales promelas*: a hierarchical approach. In progress.

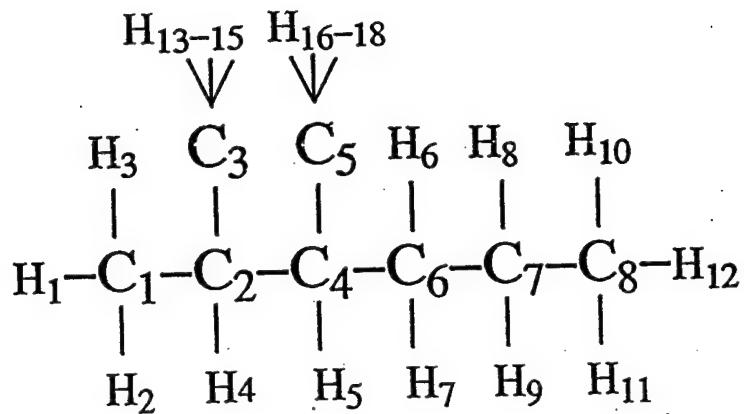
FIGURE CAPTIONS

Figure 1. Hydrogen-suppressed graph of 2,3-dimethylhexane.

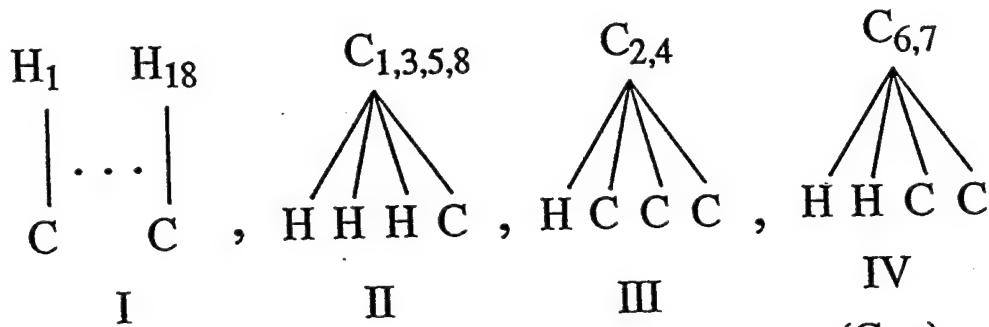
Figure 2. The calculation of IC_1 , SIC_1 , and CIC_1 , based on the first order neighborhoods for the labeled graph of 2,3-dimethylhexane.



Labeled Graph:



First Order Neighborhoods:



Subsets:

(H_{1-18})

$(C_{1,3,5,8})$

$(C_{2,4})$

$(C_{6,7})$

Probability (p_i):

$18/26$

$4/26$

$2/26$

$2/26$

$$IC_1 = - \sum p_i \cdot \log_2 p_i$$

$$= 2 \cdot 2/26 \cdot \log_2 26/2 + 4/26 \cdot \log_2 26 + 18/26 \cdot \log_2 26/18$$

$$= 1.150 \text{ bits}$$

$$SIC_1 = IC_1 / \log_2 26$$

$$= 0.353 \text{ bits}$$

$$CIC_1 = \log_2 26 - IC_1$$

$$= 2.108 \text{ bits}$$

Table I. Symbols and definitions of topological indices.

I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
\bar{I}_D^W	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
$\bar{I}C$	Information content of the distance matrix partitioned by frequency of occurrences of distance h
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
IC_r	Mean information content or complexity of a graph based on the r^{th} ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r^{th} ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r^{th} ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
h_X	Path connectivity index of order $h = 0-6$
h_{Xc}	Cluster connectivity index of order $h = 3-6$
$h_{X_{pc}}$	Path-cluster connectivity index of order $h = 4-6$
$h_{X_{ch}}$	Chain connectivity index of order $h = 3-6$
h_X^b	Bond path connectivity index of order $h = 0-6$
h_{Xc}^b	Bond cluster connectivity index of order $h = 3-6$
$h_{X_{ch}}^b$	Bond chain connectivity index of order $h = 3-6$

$h_X^p_{PC}$	Bond path-cluster connectivity index of order $h = 4-6$
h_X^v	Valence path connectivity index of order $h = 0-6$
$h_X^v_{C}$	Valence cluster connectivity index of order $h = 3-6$
$h_X^v_{Ch}$	Valence chain connectivity index of order $h = 3-6$
$h_X^v_{PC}$	Valence path-cluster connectivity index of order $h = 4-6$
P_h	Number of paths of length $h = 0-10$
J	Balaban's J index based on distance
J^B	Balaban's J index based on bond types
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii

Table II. TIs selected by variable clustering of the alkanes (octane isomers listed).

Isomer Name	H ^v	SIC ₀	SIC ₁	SIC ₄	³ X _c	⁵ X _c	P ₄	P ₈
Octane	1.288	0.173	0.218	0.477	0.000	0.000	2	0
2-methylheptane	1.233	0.173	0.248	0.561	0.342	0.000	2	0
3-methylheptane	1.228	0.173	0.248	0.598	0.254	0.000	2	0
4-methylheptane	1.215	0.173	0.248	0.503	0.254	0.000	2	0
3-ethylhexane	1.177	0.173	0.248	0.532	0.186	0.000	2	0
2,2-dimethylhexane	1.157	0.173	0.248	0.495	0.940	0.000	2	0
2,3-dimethylhexane	1.170	0.173	0.253	0.557	0.450	0.212	2	0
2,4-dimethylhexane	1.171	0.173	0.253	0.557	0.529	0.000	2	0
2,5-dimethylhexane	1.183	0.173	0.253	0.384	0.597	0.000	2	0
3,3-dimethylhexane	1.137	0.173	0.248	0.548	0.792	0.000	2	0
3,4-dimethylhexane	1.157	0.173	0.253	0.469	0.386	0.154	2	0
3-ethyl-2-methylpentane	1.096	0.173	0.253	0.490	0.405	0.154	2	0
3-ethyl-3-methylpentane	1.073	0.173	0.248	0.421	0.656	0.000	1	0
2,2,3-trimethylpentane	1.075	0.173	0.255	0.490	0.944	0.477	1	0
2,2,4-trimethylpentane	1.083	0.173	0.255	0.450	1.088	0.000	2	0
2,3,3-trimethylpentane	1.065	0.173	0.255	0.506	0.850	0.529	1	0
2,3,4-trimethylpentane	1.097	0.173	0.225	0.413	0.620	0.326	2	0
2,2,3,3-tetramethylbutane	0.997	0.173	0.218	0.218	1.253	1.179	0	0

Table III. Values of the first seven PCs for the eighteen octane isomers.

Isomer Name	PC ₁	PC ₂	PC ₃	PC ₄	PC ₅	PC ₆	PC ₇
Octane	0.328	-1.744	5.807	0.602	-0.320	-0.473	-0.433
2-methylheptane	2.181	-4.236	1.097	0.386	1.100	0.300	-0.935
3-methylheptane	2.817	-4.857	-0.307	0.921	0.368	0.366	-0.513
4-methylheptane	1.338	-2.211	0.848	-0.821	0.005	-0.541	-0.904
3-ethylhexane	1.553	-2.077	-0.348	-0.494	-0.817	-0.651	-0.290
2,2-dimethylhexane	1.163	0.007	-0.436	-0.878	1.367	1.383	0.638
2,3-dimethylhexane	2.122	-2.060	-1.546	0.502	-0.308	-0.253	-0.105
2,4-dimethylhexane	2.089	-2.306	-1.372	-0.289	-0.205	0.004	0.291
2,5-dimethylhexane	-0.769	1.340	1.473	-2.659	0.612	-0.387	-1.443
3,3-dimethylhexane	2.044	-0.573	-1.726	0.303	0.173	0.582	1.163
3,4-dimethylhexane	0.807	0.228	-0.825	-0.696	-0.730	-1.223	-0.545
3-ethyl-2-methylpentane	0.991	-0.035	-1.596	-0.672	-1.076	-1.438	0.110
3-ethyl-3-methylpentane	-0.035	2.870	-0.614	-0.909	-0.497	-1.178	0.271
2,2,3-trimethylpentane	1.136	2.191	-2.383	1.277	0.465	-0.075	0.548
2,2,4-trimethylpentane	0.377	2.377	-1.284	-1.846	0.726	0.461	1.676
2,3,3-trimethylpentane	1.318	1.825	-2.717	1.990	0.318	-0.400	0.251
2,3,4-trimethylpentane	-0.548	4.168	1.329	0.020	-1.745	-1.140	-0.039
2,2,3,3-tetramethylbutane	-4.473	12.522	2.681	4.256	1.345	-0.129	-2.627

Table IV. PC loading for the seven principal components with eigenvalues greater than 1.0 for the 74 alkanes.

Ten Most Correlated Indices										
PC	1	2	3	4	5	6	7	8	9	10
1	1^p (0.98)	W (-0.97)	1X (0.97)	l_0^w (0.97)	P_1 (0.97)	ClC_0 (0.97)	P_0 (0.97)	SiC_0 (-0.97)	6X (0.94)	IC_0 (0.94)
2	ClC_2 (0.89)	ClC_3 (0.79)	ClC_4 (0.77)	ClC_5 (0.77)	3X_C (0.76)	SiC_2 (-0.74)	4X_O (0.69)	$^4X_{PO}$ (0.69)	6X_O (0.65)	SiC_3 (-0.64)
3	SiC_1 (-0.76)	6X (0.68)	P_6 (0.67)	P_7 (0.65)	P_8 (0.55)	$^4X_{PO}$ (-0.41)	IC_1 (-0.40)	5X (0.39)	$^5X_{PO}$ (-0.39)	ClC_2 (0.38)
4	6X_C (0.64)	6X_C (0.63)	4X (-0.40)	P_4 (-0.34)	$^4X_{PC}$ (0.31)	l_{OB} (0.29)	P_7 (0.28)	ClC_5 (-0.27)	ClC_4 (-0.27)	SiC_1 (-0.24)
5	4X (-0.39)	4X_C (0.38)	$^6X_{PC}$ (-0.36)	3X_C (0.35)	P_4 (-0.34)	$^6X_{PC}$ (-0.31)	3X (-0.29)	2X (0.26)	P_3 (-0.26)	SiC_1 (0.25)
6	4X_C (0.40)	P_6 (0.39)	5X (0.37)	3X_C (0.35)	$^6X_{PC}$ (0.34)	$\overline{l_6^W}$ (-0.23)	H^p (-0.22)**	P_4 (0.19)**	IC_0 (-0.19)**	IC (-0.18)**
7	P_8 (0.59)	P_7 (0.38)	4X_O (0.30)	6X_C (-0.29)	P_5 (-0.20)**	5X_O (-0.19)**	5X (-0.19)**	3X_O (0.17)**	6X (-0.17)**	O (0.16)**

ns Not significant at the $p \leq 0.05$ level.

Table V. PC loading for the 10 principal components with eigenvalues greater than 1.0 for the 219 STARLIST chemicals.

Ten Most Correlated Indices

PC	1	2	3	4	5	6	7	8	9	10
1	P_o (0.97)	1X (0.96)	9X (0.96)	P_1 (0.96)	3X (0.95)	W (0.95)	$^4X^p$ (0.95)	$^4X^p$ (0.95)	M_2 (0.95)	M_1 (0.94)
2	SIC_4 (-0.86)	SIC_3 (-0.86)	SIC_6 (-0.86)	SIC_8 (-0.86)	SIC_6 (0.80)	SIC_8 (0.80)	SIC_4 (0.80)	SIC_2 (-0.78)	SIC_3 (0.76)	IC_2 (-0.74)
3	IC_2 (-0.67)	SIC_1 (0.65)	$^5X_{Ch}$ (0.63)	IC_1 (-0.63)	5X_C (0.61)	SIC_6 (0.61)	6X_C (0.60)	$^6X_{Ch}$ (0.59)	IC_3 (-0.58)	
4	J (0.83)	J^x (0.73)	J^y (0.73)	J^z (0.62)	$^3X^p_C$ (0.56)	3X_C (0.55)	$^6X_{Ch}$ (-0.44)	P_{10} (-0.42)	$^5X^p_{Ch}$ (-0.41)	$^6X_{Ch}$ (-0.41)
5	IC_6 (-0.45)	SIC_6 (-0.43)	J^x (0.36)	J (0.35)	$^4X^p_{Ch}$ (0.35)	IC_6 (0.35)	SIC_1 (-0.34)	$^6X^p_C$ (-0.33)	$^6X_{Ch}$ (0.33)	
6	$^4X^p_o$ (0.57)	4X_C (0.57)	P_o (0.48)	P_g (0.46)	$^4X^p_C$ (0.44)	P_{10} (0.42)	$^3X^p_o$ (0.35)	3X_C (0.33)	$^5X^p_C$ (-0.32)	P_7 (0.31)
7	6X_C (-0.43)	$^6X^p_o$ (-0.42)	$^5X^p_o$ (-0.42)	$^3X_{Ch}$ (0.40)	5X_C (-0.39)	$^4X_{Ch}$ (0.31)	$^4X^p_{Ch}$ (0.29)	$^4X^p_{Po}$ (-0.26)	$^5X_{Ch}$ (0.26)	$^5X^p_{Ch}$ (0.21)
8	$^4X^p_C$ (0.49)	$^3X^p_C$ (0.40)	$^2X^p$ (0.29)	$^6X^p_C$ (-0.27)	6X_C (-0.26)	J^y (-0.24)	$^1X^y$ (0.23)	$^9X^y$ (0.22)	J^B (-0.21)	P_g (-0.20)
9	$^3X_{Ch}$ (0.73)	$^4X_{Ch}$ (0.47)	$^4X^p_{Ch}$ (0.43)	$^6X^p_{Ch}$ (-0.21)	$^5X^p_{Ch}$ (-0.21)	6X_C (-0.19)	$^6X_{Ch}$ (-0.16)	J^x (-0.16)	$^6X^p_{Ch}$ (-0.15)	
10	IC_6 (0.35)	H^y (0.25)	J^x (-0.24)	J^y (0.24)	SIC_6 (0.21)	I_{ORB} (-0.21)	$^6X^p_C$ (-0.21)	J^y (-0.20)	J^B (-0.19)	$^1X^p$ (0.19)

THE RELATIVE EFFECTIVENESS OF TOPOLOGICAL, GEOMETRICAL AND
QUANTUM CHEMICAL PARAMETERS IN ESTIMATING MUTAGENICITY OF
CHEMICALS

Subhash C. Basak*

Brian D. Gute

and

Gregory D. Grunwald

Natural Resources Research Institute

University of Minnesota

5013 Miller Trunk Highway

Duluth, MN 55811, USA

Proceedings of the 7th International Workshop on QSAR in Environmental Sciences, SETAC Press, in press, 1997.

* To whom all correspondence may be addressed.

Abstract - Adequate experimental data necessary for hazard assessment is not available for the majority of environmental pollutants and chemicals in commerce. This has led to the increasing use of theoretical structural parameters in the hazard estimation of such chemicals. In this paper we have used a hierarchical QSAR approach involving topological indices, geometrical (3-D) indices, and quantum chemical indices to estimate the mutagenicity of a set of ninety-five aromatic and heteroaromatic amines. The results show that topological indices explain the major part of the variance in mutagenicity. The addition of quantum chemical indices to the set of descriptors makes some improvement in the predictive models.

Keywords - Topological Indices Geometrical Parameters Quantum Chemical Parameters
Mutagenicity Hierarchical QSAR

INTRODUCTION

The assessment of the environmental and human health hazard posed by chemicals is frequently carried out using insufficient experimental data. This is true for industrial chemicals, as well as for substances identified in industrial effluent, hazardous waste sites and environmental monitoring surveys (Auer *et al.* 1990). In 1984, the National Research Council (NRC) studied the availability of toxicity data on industrial chemicals and found that many of these chemicals have very little or no test data (1984). About 15 million distinct chemical entities have been registered with the Chemical Abstract Service (CAS) and the list is growing by nearly 750,000 per year. Out of these chemicals, about 1,000 enter into societal use every year (Arcos 1987). Very few of these chemicals have empirical properties needed for hazard assessment. In the United States, the Toxic Substances Control Act (TSCA) inventory has over 72,000 entries and the list is growing by nearly 3,000 per year (GAO 1993). Of the some 3,000 chemicals submitted yearly to the United States Environmental Protection Agency (USEPA) for the premanufacture notification (PMN) process, less than 50% have any experimental data at all, less than 15% have empirical mutagenicity data, and only about 6% have ecotoxicological and environmental fate data. The Superfund list of hazardous substances has only limited data for many of the over 700 chemicals as well (Auer *et al.* 1990).

This pervasive lack of empirical data shows the real need for the development of methods which can estimate environmental and toxic properties of chemicals using parameters which can be calculated directly from molecular structure. In recent years we have been involved in the development of such models (Basak and Magnuson 1983; Basak 1987, 1990; Basak *et al.* 1988, 1994; Balaban *et al.* 1994; Basak and Grunwald 1994a, 1994b, 1995a-1995e, 1996; Basak,

Bertelsen and others 1995; Basak, Gute and others 1995, 1996a-1996c; Basak, Grunwald and others 1996; Basak and Gute 1996). Specifically, we have used graph theoretic indices, geometrical (3-D) parameters, and semi-empirical quantum chemical indices in the development of quantitative structure-activity relationship (QSAR) models pertinent to biomedical chemistry and toxicology. In this paper we have used a hierarchical approach in the development of QSARs for a group of ninety-five aromatic and heteroaromatic amines using topological indices, 3-D parameters and a set of quantum chemical descriptors.

The purpose in using a hierarchical approach is to begin to look at the importance of the contribution of different classes of parameters to modeling physicochemical or biologically relevant properties. To this end we ask the question, what non-empirical molecular information is adequate for the estimation of mutagenic potency? Is specific chemical or quantum chemical information necessary or do simple structural descriptors do an adequate job? These questions should lead us to a deeper understanding of the principles and molecular basis for determining mutagenic potency.

THEORETICAL METHODS

Database

A set of 95 aromatic and heteroaromatic amines, previously collected from the literature by Debnath *et al.* (1992), were used to study mutagenic potency. The mutagenic activities of these compounds in *S. typhimurium* TA98 + S9 microsomal preparation are expressed as the mutation rate, $\ln(R)$, in natural logarithm (revertants/nanomole). Table I lists the compounds used

in this study and their experimentally measured mutation rates.

Computation of Topological Indices

Topological indices used in this study have been calculated by POLLY 2.3 (Basak *et al.* 1988) which can calculate a total of 102 indices. These indices include Wiener index (Wiener 1947), connectivity indices (Kier and Hall 1986; Randic 1975), information theoretic indices defined on distance matrices of graphs (Raychaudhury *et al.* 1984; Bonchev and Trinajstic 1977), a set of parameters derived on the neighborhood complexity of vertices in hydrogen-filled molecular graphs (Basak 1987; Basak and Magnuson 1983; Basak *et al.* 1980; Roy *et al.* 1984), as well as Balaban's J indices (Balaban 1982, 1983, 1986). Table II provides brief definitions for the topological indices included in this study.

Computation of Geometrical Indices

van der Waal's volume, V_w , (Bondi 1964; Moriguchi *et al.* 1975; Moriguchi and Kanada 1977) was calculated using *Sybyl 6.2* from Tripos Associates, Inc (1994). The 3-D Wiener numbers (Bogdanov *et al.* 1989) were calculated by *Sybyl* using an SPL (Sybyl Programming Language) program developed in our laboratory. Calculation of 3-D Wiener numbers consists of the sum entries in the upper triangular submatrix of the topographic Euclidean distance matrix for a molecule. The 3-D coordinates for the atoms were determined using *CONCORD 3.2.1* (Tripos 1993). Two variants of the 3-D Wiener number were calculated: ${}^{3D}W_H$ and ${}^{3D}W$. For ${}^{3D}W_H$,

hydrogen atoms are included in the computations and for ^{3D}W , hydrogen atoms are excluded from the computations.

Computation of Quantum Chemical Parameters

The quantum chemical parameters E_{HOMO} , E_{HOMO1} , E_{LUMO} , E_{LUMO1} , ΔH_f , and μ were calculated for all of the following semi-empirical Hamiltonians: AM1, PM3, MNDO, MINDO/3. These parameters were calculated by *MOPAC 6.00* in the *SYBYL* interface (Stewart 1985). One difficulty was encountered in using the MINDO/3 Hamiltonian. This particular interface does not include the information necessary for handling bromine, present in three of the ninety-five molecules. To avoid omitting any compounds from one of the models, we accounted for the bromine by substituting dummy atoms which were assigned the Gasteiger-Huckel charges calculated for the original bromine atoms. These molecules containing the dummy atoms with assigned charges were then entered into *MOPAC* for calculation.

Data Reduction

Initially, all TIs were transformed by the natural logarithm of the index plus one. This was done since the scale of some indices may be several orders of magnitude greater than that of other indices and other indices may equal zero. The geometric indices were transformed by the natural logarithm of the index for consistency, the addition of one was unnecessary.

The set of 91 TIs was partitioned into two distinct sets: topostructural indices and

topochemical indices. Topostructural indices are indices which encode information about the adjacency and distances of atoms (vertices) in molecular structures (graphs) irrespective of the chemical nature of the atoms involved in the bonding or factors like hybridization states of atoms and number of core/valence electrons in individual atoms. Topochemical indices are parameters which quantify information regarding the topology (connectivity of atoms) as well as specific chemical properties of the atoms comprising a molecule. Topochemical indices are derived from weighted molecular graphs where each vertex (atom) is properly weighted with selected chemical/physical properties. These sets of the indices are shown in Table III.

According to Topliss and Edwards, in conducting QSAR studies it is important to bear in mind that the indiscriminate use of too many independent variables can lead to spurious (chance) correlations (1979). Using their findings, we have determined that for a set of 95 compounds no more than 60 independent variables can be used in generating regression analyses with explained variance (R^2) of 0.7 or greater. It must be kept in mind that this is the total number of variables initially used in modeling, not the final number of variables used in the model. This number of independent variables should keep the probability of chance correlations below the 0.01 level.

To reduce the number of independent variables that we would use for model construction, the sets of topostructural and topochemical indices were further divided into subsets, or clusters, based on the correlation matrix using the SAS procedure VARCLUS (SAS 1988). The VARCLUS procedure divides the set of indices into disjoint clusters so that each cluster is essentially unidimensional.

From each cluster we selected the index most correlated with the cluster, as well as any indices which were poorly correlated with the cluster ($r < 0.70$). These indices were then used in

the modeling of mutagenic potency of aromatic and heteroaromatic amines. The variable clustering and selection of indices was performed independently for both the topostructural and topochemical subsets.

Statistical Analysis and Hierarchical QSAR

Regression modeling was accomplished using the SAS procedure REG on thirteen sets of indices. These sets were constructed as part of a hierarchical approach to QSAR model development. The hierarchy begins with the simplest indices, the topostructural. After using the topostructural indices to model the activity, we then proceed to add the next level of complexity, the topochemical indices from the clustering procedure, and proceed to model the activity using these parameters. Likewise, the indices included in the model selected from this procedure are combined with the indices from the next level, the geometrical indices, and modeling is conducted once again. Finally, the best model utilizing topostructural, topochemical and geometrical indices is combined with the quantum chemical parameters and modeling is conducted. This final step was repeated four times, each time using quantum chemical parameters from a different semi-empirical Hamiltonian, namely, AM1, PM3, MNDO, MINDO/3. Thus quantum chemical models are developed individually, one using the AM1 parameters, one using the MNDO parameters, one using the PM3 parameters, and one using the MINDO/3 parameters. The regression analysis resulted in the final selection of indices for each of the models.

RESULTS AND DISCUSSION

The variable clustering of topostructural and topochemical indices resulted in 8 topostructural and 13 topochemical indices being retained for model construction (see Table III). The results for the all possible subsets regression analyses have been summarized in Table IV. Since all sets were well under 25 parameters, all possible subsets regression was used for all analyses.

As can be seen from Table IV, using only the topostructural class of indices resulted in a four parameter model to estimate $\ln(R)$ with a variance explained (R^2) of 72.1% and a standard error (s) of 1.04 (equation 1). The P_0 and J indices are related to the size and shape of molecular graphs; the ${}^4\chi_{PC}$ encodes information about the degree of branching of molecular graphs; the O parameter is related to the degree of symmetry of graphs (Basak *et al.* 1987). Therefore, size, branching, and symmetry (or complexity) of skeletal graphs corresponding to molecular structures seem to be the predominant factors in determining mutagenic potency of the set of 95 aromatic amines.

The second step of the hierarchical method combined the four topostructural parameters from equation 1 with the set of thirteen topochemical parameters. The resulting model for estimation of $\ln(R)$ included six parameters (equation 8) which had an R^2 of 75.2% and a s of 0.99. Thus we see that the addition of topochemical information does lead to an increase in the explained variance, improving our model without greatly increasing the number of independent variables. The independent variables of equation 8 quantify : a) shape and size of molecular graphs (J , P_0), b) branching (${}^4\chi_{PC}$), c) molecular complexity / redundancy (SIC_2 , SIC_4), and d)

degree of cyclicity (${}^5\chi^b c$). It may be mentioned that we have found very similar set of topostructural and topochemical parameters useful in estimating normal boiling point, octanol water partition coefficient (Basak, Gute and others 1996c), and vapor pressure (Basak, Gute and others 1996d) of diverse sets of molecules.

The next step of the hierarchical method takes this topostructural + topochemical model and adds the three geometric indices, however, this actually led to a decrease in the explained variance. As part of model construction, it became necessary to eliminate P_0 from the set of indices when adding the hydrogen-suppressed 3-D Wiener number because of resulting problems with variance inflation between the two parameters. As a result, the model which retained the geometric parameter had a slightly lower R^2 and s values than the model using topostructural and topochemical only (equation 9). This being the case, we chose to use the parameters from equation 8 in the following modeling with the quantum chemical parameters. Thus, the last four models were all constructed with the six parameters from equation 8 and all six quantum chemical parameters for the particular Hamiltonian methodology available for modeling.

As can be seen from Table IV, the AM1 parameters made the most significant contribution to our hierarchical modeling procedure ($R^2 = 79.1\%$, $s = 0.92$). The other three methods showed only minimal improvement over the topostructural + topochemical model.

Finally, individual models using only topochemical, only geometrical, and only quantum chemical parameters were constructed to further our understanding of the individual contribution of these different types of parameters. The topochemical model was the strongest of the three, with the geometrical and quantum chemical models showing little effectiveness. The topochemical model included six parameters and did show a slight increase in explained variance and standard

error over the topostructural model.

The goal of the paper was to investigate the relative effectiveness of theoretical structural parameters; namely topostructural, topochemical, geometrical and quantum chemical parameters; in predicting the mutagenicity of a set of aromatic and heteroaromatic amines. To this end, we used a hierarchical approach in the development of QSARs using four classes of molecular descriptors.

The results show that the topostructural parameters explain a large fraction of the variance (R^2) in the mutagenic potency of the amines. The best model in this area explained about 72% of variance in mutagenicity using O , $^4\chi_{PC}$, P_0 , J . These indices do not contain any explicit chemical information about the molecules. The large explained variance probably indicates that general structural features like size, shape, symmetry, and branching play a major role in determining mutagenic potency. The addition of topochemical variables made some improvement in the explained variance. The best model using topostructural and topochemical indices explained about 75% of variance in mutagenicity. The addition of geometrical parameters, however, did not make any improvement in estimation. Finally, the addition of quantum chemical parameters was attempted. Indices from AM1, PM3, MNDO and MINDO3 were used separately in developing the QSAR models. While addition of the heat of formation, dipole moment and E_{HOMO1} parameters calculated by the AM1 method provided some improvement in the estimation of $\ln(R)$, parameters calculated by PM3, MINDO3 and MNDO did not make any significant improvement in the estimation of mutagenic potency. The calculated values for the parameters used in the hierarchical model which included the AM1 parameters (equation 10) are presented in Table V. These values represent the original, non-transformed values for all indices used in equation 10.

Using the same set of aromatic amines Debnath *et al.* (1992) developed various QSAR models using hydrophobicity ($\log P$, octanol/water), E_{HOMO} and E_{LUMO} calculated by the AM1 Hamiltonian and some indicator variables. For the largest subset ($n = 88$), they derived the following model:

$$\ln (R) = 7.20 + 1.08(\log P) + 1.28(E_{HOMO}) - 0.73(E_{LUMO}) + 1.46(I_L)$$

$$s = 0.860, F = 12.6, R^2 = 0.806$$

The model in equation 10 is comparable to the model developed by Debnath *et al.* and uses all the 95 aromatic amines as compared to a smaller subset ($n=88$) used in their study.

ACKNOWLEDGMENTS

This is contribution number 197 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported in part by grants F49620-94-1-0401 and F49620-96-1-0330 from the United States Air Force, a grant from Exxon Corporation and the Structure-activity Relationship Consortium (SARCON) of the Natural Resources Research Institute of the University of Minnesota.

REFERENCES

Arcos JC. 1987. Structure-activity relationships: criteria for predicting the carcinogenic activity of chemical compounds. *Environ Sci Tech.* 21:743-745.

Auer CM, Nabholz JV, Baetcke KP. 1990. Mode of action and the assessment of chemical hazards in the presence of limited data: use of structure-activity relationships (SAR) under TSCA, section 5. *Environ Health Perspect.* 87:183-197.

Balaban AT. 1982. Highly discriminating distance-based topological index. *Chem Phys Lett.* 89:399-404.

Balaban AT. 1983. Topological indices based on topological distances in molecular graphs. *Pure and Appl Chem.* 55:199-206.

Balaban AT. 1986. Chemical graphs. Part 48. Topological index J for heteroatom-containing molecules taking into account periodicities of element properties. *Math Chem (MATCH).* 21:115-122.

Balaban AT, Basak SC, Colburn T, Grunwald G. 1994. Correlation between structure and normal boiling points of haloalkanes C1-C4 using neural networks. *J Chem Inf Comput Sci.* 34:1118-1121.

Basak SC. 1987. Use of molecular complexity indices in predictive pharmacology and toxicology: a QSAR approach. *Med Sci Res.* 15:605-609.

Basak SC. 1990. A nonempirical approach to predicting molecular properties using graph-theoretic invariants. In: Karcher W, Devillers J, editors. *Practical applications of quantitative structure-activity relationships (QSAR) in environmental chemistry and toxicology.* Dordrecht/Boston/London: Kluwer Academic Publishers. p 83-103.

Basak SC, Bertelsen S, Grunwald G. 1994. Application of graph theoretical parameters in quantifying molecular similarity and structure-activity studies. *J Chem Inf Comput Sci.* 34:270-276.

Basak SC, Bertelsen S, Grunwald GD. 1995. Use of graph theoretic parameters in risk assessment of chemicals. *Toxicology Letters.* 79:239-250.

Basak SC, Grunwald GD. In press 1994a. Use of topological space and property space in selecting structural analogs. *Mathematical Modelling and Scientific Computing.*

Basak SC, Grunwald GD. 1994b. Molecular similarity and risk assessment: analog selection and property estimation using graph invariants. *SAR and QSAR in Environmental Research.* 2:289-307.

Basak SC, Grunwald GD. 1995a. Estimation of lipophilicity from molecular structural similarity. *New Journal of Chemistry.* 19:231-237.

Basak SC, Grunwald GD. 1995b. Predicting genotoxicity of chemicals using nonempirical parameters. In: Rao RS, Deo MG, Sanghui LD, editors. Proceeding of the XVI International Cancer Congress. Bologna, Italy: Monduzzi. p 413-416.

Basak SC, Grunwald GD. 1995c. Molecular similarity and estimation of molecular properties. *J Chem Inf Comput Sci.* 35:366-372.

Basak SC, Grunwald GD. 1995d. Tolerance space and molecular similarity. *SAR and QSAR in Environmental Research.* 3:265-277.

Basak SC, Grunwald GD. 1995e. Predicting mutagenicity of chemicals using topological and quantum chemical parameters: a similarity based study. *Chemosphere.* 31:2529-2546.

Basak SC, Grunwald GD. In preparation 1996. Characterization of relative proximity of molecules in structure space: development of a molecular ruler using octane isomers. *Mathl Modelling Sci Computing.*

Basak SC, Grunwald GD, Niemi GJ. In press 1996. Use of graph theoretical and geometrical molecular descriptors in structure-activity relationships. In: Balaban AT, editor. From chemical topology to three dimensional molecular geometry. Plenum Press.

Basak SC, Gute BD. In press 1996. Use of graph theoretic parameters in predicting inhibition of microsomal hydroxylation of anilines by alcohols: a molecular similarity approach. Proceedings of the international congress on hazardous waste: impact on human and ecological health. Atlanta GA.

Basak SC, Gute BD, Drewes LR. 1996a. Predicting blood-brain transport of drugs: a computational approach. *Pharm Res.* 13:775-778.

Basak SC, Gute BD, Grunwald GD. In press 1995. Development and applications of molecular similarity methods using nonempirical parameters. *Mathl Modelling Sci Computing.*

Basak SC, Gute BD, Grunwald GD. 1996b. Estimation of normal boiling points of haloalkanes using molecular similarity. *Croat Chim Acta.* 69: In press.

Basak SC, Gute BD, Grunwald GD. 1996c. A comparative study of topological and geometrical parameters in estimating normal boiling point and octanol-water partition coefficient. *J Chem Inf Comput Sci.* 36:1054-1060

Basak SC, Gute BD, Grunwald GD. 1996d. Use of topostructural, topochemical and geometrical parameters in the prediction of vapor pressure: a hierarchical QSAR approach, *J Chem Inf Comput Sci*, submitted.

Basak SC, Hartiss DK, Magnuson VR. 1988. POLLY 2.3: Copyright of the University of Minnesota.

Basak SC, Magnuson VR. 1983. Molecular topology and narcosis: a quantitative structure-

activity relationship (QSAR) study of alcohols using complementary information content (CIC). *Arzneim Forsch.* 33:501-503.

Basak SC, Magnuson VR, Niemi GJ, Regal RR, Veith GD. 1987. Topological indices: their nature, mutual relatedness, and applications. *Mathematical Modelling*, 8: 300-305.

Basak SC, Roy AB, Ghosh JJ. 1980. In: Avula XJR, Bellman R, Luke YL, Rigler AK, editors. *Proceedings of the Second International Conference on Mathematical Modelling*. University of Missouri - Rolla; Rolla, MO. Vol. II, p 851.

Bogdanov B, Nikolic S, Trinajstic N. 1989. On the three-dimensional Wiener number. *J Math Chem.* 3:299-309.

Bonchev D, Trinajstic N. 1977. Information theory, distance matrix and molecular branching. *J Chem Phys.* 67:4517-4533.

Bondi A. 1964. van der Waal's volumes and radii. *J Phys Chem.* 68:441-451.

Debnath AK, Debnath G, Shusterman AJ, Hansch C. 1992. A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the ames test: 1. Mutagenicity of aromatic and heteroaromatic amines in *Salmonella typhimurium* TA98 and TA100. *Environ Mol Mutagen.* 19:37-52.

GAO. 1993. EPA toxic substances program: long-standing information planning problems must be addressed. US General Accounting Office, Accounting and Information Management Division. GAO/AIMD-94-25. Washington DC: US GAO.

Kier LB, Hall LH. 1986. Molecular connectivity in structure-activity analysis. Letchworth, Hertfordshire, UK: Research Studies Press. 262 p.

Moriguchi I, Kanada Y. 1977. Use of van der Waal's volume in structure-activity studies. *Chem Pharm Bull.* 25:926-935.

Moriguchi I, Kanada Y, Komatsu K. 1976. van der Waal's volume and the related parameters for hydrophobicity in structure-activity studies. *Chem Pharm Bull.* 24:1799-1806.

NRC. 1984. Toxicity testing: strategies to determine needs and priorities. Washington DC: National Academy Press. 84 p.

Randic, M. 1975. On characterization of molecular branching. *J Am Chem Soc.* 97:6609-6615.

Raychaudhury C, Ray SK, Ghosh JJ, Roy AB, Basak SC. 1984. Discrimination of isomeric structures using information theoretic topological indices. *J Comput Chem.* 5:581-588.

Roy AB, Basak SC, Harris DK, Magnuson VR. 1984. In: Avula XJR, Kalman RE, Liapis AI, Rodin EY, editors. *Mathematical Modelling in Science and Technology*. Pergamon Press: New York, p745.

SAS Institute Inc. In *SAS/STAT User's Guide, Release 6.03 Edition*. SAS Institute Inc.: Cary, NC, 1988; Chapters 28 and 34, pp. 773-875, 949-965.

Stewart JJP. 1990. MOPAC Version 6.00. QCPE #455. Frank J Seiler Research Laboratory: US Air Force Academy CO.

Topliss JG, Edwards RP. 1979. Chance factor in studies of quantitative-structure activity relationships. *J Med Chem.* 22:1238-1244.

Tripos Associates, Inc. *SYBYL Version 6.2*. Tripos Associates, Inc.: St. Louis, MO, 1994.

Tripos Associates, Inc. *CONCORD Version 3.2.1*. Tripos Associates, Inc.: St. Louis, MO, 1993.

Wiener H. 1947. Structural determination of paraffin boiling points. *J Am Chem Soc.* 69:17-20.

LEGEND FOR FIGURE:

Figure 1. Scatterplot for observed $\ln(R)$ vs. estimated $\ln(R)$ using equation 10 for a set of 95 aromatic and heteroaromatic amines.

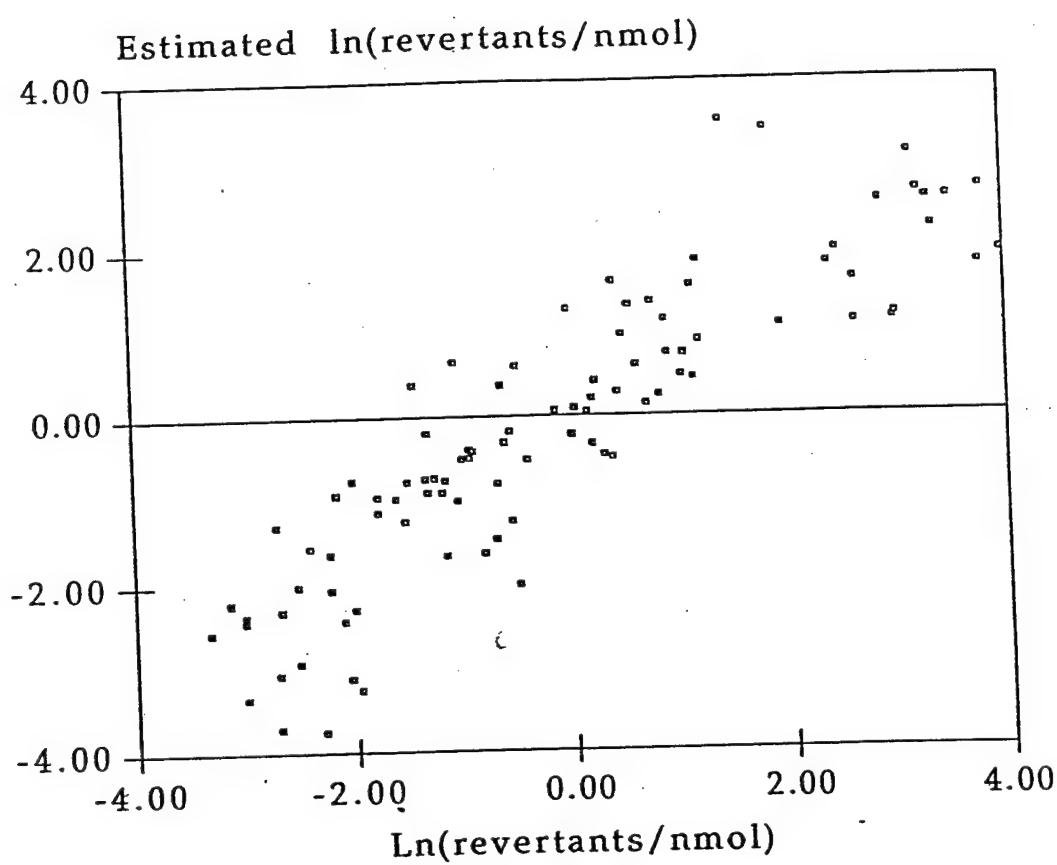


Table I. Observed and estimated mutagenic potency [$\ln(\text{revertants/nmol})$] for ninety-five aromatic and heteroaromatic amines.

No.	Compound	Exp. $\ln(R)$	Est. $\ln(R)$ (eq.10)
1	2-bromo-7-aminofluorene	2.62	1.10
2	2-methoxy-5-methylaniline (p-cresidine)	-2.05	-3.13
3	5-aminoquinoline	-2.00	-2.30
4	4-ethoxyaniline (p-phenetidine)	-2.30	-3.76
5	1-aminonaphthalene	-0.60	-0.32
6	4-aminofluorene	1.13	0.44
7	2-aminoanthracene	2.62	1.61
8	7-aminofluoranthene	2.88	2.54
9	8-aminoquinoline	-1.14	-1.66
10	1,7-diaminophenazine	0.75	1.36
11	2-aminonaphthalene	-0.67	-0.80
12	4-aminopyrene	3.16	3.10
13	3-amino-3'-nitrobiphenyl	-0.55	-0.19
14	2,4,5-trimethylaniline	-1.32	-0.74
15	3-aminofluorene	0.89	0.74
16	3,3'-dichlorobenzidine	0.81	0.24
17	2,4-dimethylaniline (2,4-xylidine)	-2.22	-1.63
18	2,7-diaminofluorene	0.48	0.97
19	3-aminofluoranthene	3.31	2.57
20	2-aminofluorene	1.93	1.08
21	2-amino-4'-nitrobiphenyl	-0.62	0.37
22	4-aminobiphenyl	-0.14	0.06
23	3-methoxy-4-methylaniline (o-cresidine)	-1.96	-3.27

24	2-aminocarbazole	0.60	0.60
25	2-amino-5-nitrophenol	-2.52	-2.01
26	2,2'-diaminobiphenyl	-1.52	-1.24
27	2-hydroxy-7-aminofluorene	0.41	1.61
28	1-aminophenanthrene	2.38	1.80
29	2,5-dimethylaniline (2,5-xylidine)	-2.40	-1.55
30	4-amino-2'-nitrobiphenyl	-0.92	-0.50
31	2-amino-4-methylphenol	-2.10	-2.43
32	2-aminophenazine	0.55	1.32
33	4-aminophenylsulfide	0.31	-0.47
34	2,4-dinitroaniline	-2.00	-0.75
35	2,4-diaminoisopropylbenzene	-3.00	-3.36
36	2,4-difluoroaniline	-2.70	-1.29
37	4,4'-methylenedianiline	-1.60	-0.97
38	3,3'-dimethylbenzidine	0.01	-0.23
39	2-aminofluoranthene	3.23	2.66
40	2-amino-3'-nitrobiphenyl	-0.89	-0.42
41	1-aminofluoranthene	3.35	2.23
42	4,4'-ethylenebis (aniline)	-2.15	-0.92
43	4-chloroaniline	-2.52	-2.94
44	2-aminophenanthrene	2.46	1.96
45	4-fluoroaniline	-3.32	-2.57
46	9-aminophenanthrene	2.98	1.13
47	3,3'-diaminobiphenyl	-1.30	-0.20
48	2-aminopyrene	3.50	2.58
49	2,6-dichloro-1,4-phenylenediamine	-0.69	-1.46

50	2-amino-7-acetamidofluorene	1.18	0.89
51	2,8-diaminophenazine	1.12	1.55
52	6-aminoquinoline	-2.67	-2.31
53	4-methoxy-2-methylaniline (m-Cresidine)	-3.00	-2.44
54	3-amino-2'-nitrobiphenyl	-1.30	-0.90
55	2,4'-diaminobiphenyl	-0.92	-0.40
56	1,6-diaminophenazine	0.20	0.20
57	4-aminophenyldisulfide	-1.03	-1.00
58	2-bromo-4,6-dinitroaniline	-0.54	-1.25
59	2,4-diamino-n-butylbenzene	-2.70	-3.72
60	4-aminophenylether	-1.14	-0.76
61	2-aminobiphenyl	-1.49	-0.77
62	1,9-diaminophenazine	0.04	0.09
63	1-aminofluorene	0.43	0.28
64	8-aminofluoranthene	3.80	2.69
65	2-chloroaniline	-3.00	-2.37
66	2-amino- α,α,α -trifluorotoluene	-0.80	-1.63
67	2-amino-1-nitronaphthalene	-1.17	-0.90
68	3-amino-4'-nitrobiphenyl	0.69	0.14
69	4-bromoaniline	-2.70	-3.08
70	2-amino-4-chlorophenol	-3.00	-2.39
71	3,3'-dimethoxybenzidine	0.15	0.05
72	4-cyclohexylaniline	-1.24	-0.73
73	4-phenoxyaniline	0.38	-0.50
74	4,4'-methylenebis (o-ethylaniline)	-0.99	-0.51
75	2-amino-7-nitrofluorene	3.00	1.19

76	benzidine	-0.39	-0.52
77	1-amino-4-nitronaphthalene	-1.77	-0.95
78	4-amino-3'-nitrobiphenyl	1.02	0.47
79	4-amino-4'-nitrobiphenyl	1.04	0.73
80	1-aminophenazine	-0.01	1.28
81	4,4'-methylenebis (o-fluoroaniline)	0.23	0.41
82	4-chloro-2-nitroaniline	-2.22	-2.06
83	3-aminoquinoline	-3.14	-2.22
84	3-aminocarbazole	-0.48	0.60
85	4-chloro-1,2-phenylenediamine	-0.49	-2.01
86	3-aminophenanthrene	3.77	1.79
87	3,4'-diaminobiphenyl	0.20	-0.34
88	1-aminoanthracene	1.18	1.86
89	1-aminocarbazole	-1.04	0.65
90	9-aminoanthracene	0.87	1.15
91	4-aminocarbazole	-1.42	0.38
92	6-aminochrysene	1.83	3.41
93	1-aminopyrene	1.43	3.51
94	4-4'-methylenebis (o-isopropyl-aniline)	-1.77	-1.13
95	2,7-diaminophenazine	3.97	1.93

Table II. Symbols and definitions of topological and geometrical parameters.

I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
\overline{I}_D^W	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
\overline{IC}	Information content of the distance matrix partitioned by frequency of occurrences of distance h
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
IC_r	Mean information content or complexity of a graph based on the r^{th} ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r^{th} ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r^{th} ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
$^h\chi$	Path connectivity index of order $h = 0-6$
$^h\chi_C$	Cluster connectivity index of order $h = 3-5$
$^h\chi_{PC}$	Path-cluster connectivity index of order $h = 4-6$
$^h\chi_{Ch}$	Chain connectivity index of order $h = 5, 6$
$^h\chi^b$	Bond path connectivity index of order $h = 0-6$
$^h\chi_C^b$	Bond cluster connectivity index of order $h = 3, 5$

${}^h\chi_{Ch}^b$	Bond chain connectivity index of order $h = 5, 6$
${}^h\chi_{PC}^b$	Bond path-cluster connectivity index of order $h = 4-6$
${}^h\chi^v$	Valence path connectivity index of order $h = 0-6$
${}^h\chi_C^v$	Valence cluster connectivity index of order $h = 3, 5$
${}^h\chi_{Ch}^v$	Valence chain connectivity index of order $h = 5, 6$
${}^h\chi_{PC}^v$	Valence path-cluster connectivity index of order $h = 4-6$
P_h	Number of paths of length $h = 0-10$
J	Balaban's J index based on distance
J^B	Balaban's J index based on bond types
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii
V_w	van der Waal's volume
${}^{3D}W$	3-D Wiener number for the hydrogen-suppressed geometric distance matrix
${}^{3D}W_H$	3-D Wiener number for the hydrogen-filled geometric distance matrix

Table III. Classification of parameters used in developing models for mutagenic potency (lnR).

Topological	Topochemical	Geometric	Quantum Chemical: AM1, PM3, MNDO, MINDO/3
I_D^W	I_{ORB}	V_w	E_{HOMO}
\overline{I}_D^W	$IC_0 - IC_6$	${}^{3D}W$	E_{HOMOI}
W	$SIC_0 - SIC_6$	${}^{3D}W_H$	E_{LUMO}
I^D	$CIC_0 - CIC_6$		E_{LUMOI}
H^V	${}^0\chi^b - {}^6\chi^b$		ΔH^f
H^D	${}^3\chi^b_C$ and ${}^5\chi^b_C$		μ
$\overline{I}C$	${}^5\chi^b_{Ch}$ and ${}^6\chi^b_{Ch}$		
O	${}^4\chi^b_{PC} - {}^6\chi^b_{PC}$		
M_1	${}^0\chi^v - {}^6\chi^v$		
M_2	${}^3\chi^v_C$ and ${}^5\chi^v_C$		
${}^0\chi - {}^6\chi$	${}^5\chi^b_{Ch}$ and ${}^6\chi^b_{Ch}$		
${}^3\chi_C$ and ${}^5\chi_C$	${}^4\chi^b_{PC} - {}^6\chi^b_{PC}$		
${}^5\chi_{Ch}$ and ${}^6\chi_{Ch}$	J^B		
${}^4\chi_{PC} - {}^6\chi_{PC}$	J^X		
$P_0 - P_{10}$	J^Y		
J			

Table IV. Summary of regression results for all classes of parameters.

eq.	parameter class	variables included	F	R^2	s
1	topostructural	$O, {}^4\chi_{PC}, P_0, J$	58.1	0.721	1.04
2	topochemical	$IC_4, SIC_2, SIC_4, {}^4\chi^v, {}^5\chi^b_c, {}^4\chi^b_{PC}$	41.1	0.737	1.02
3	geometric	${}^{3D}W$	61.8	0.399	1.50
4	Q_C : AM1	E_{HOMO1}, E_{LUMO}, μ	31.8	0.512	1.37
5	Q_C : MNDO	E_{HOMO1}, E_{LUMO}	54.7	0.543	1.31
6	Q_C : MINDO/3	$E_{HOMO}, E_{LUMO}, \Delta H_f$	32.4	0.517	1.36
7	Q_C : PM3	$E_{HOMO}, E_{HOMO1}, E_{LUMO}$	30.0	0.497	1.39
8	topostructural + topochemical	${}^4\chi_{PC}, P_0, J, SIC_2, SIC_4, {}^5\chi^b_c$	44.5	0.752	0.99
9	topostructural + topochemical + geometric	${}^4\chi_{PC}, J, SIC_2, SIC_4, {}^5\chi^b_c, {}^{3D}W$	42.9	0.746	1.00
10	topostructural + topochemical + geometric + AM1	${}^4\chi_{PC}, P_0, J, SIC_2, SIC_4, {}^5\chi^b_c, E_{HOMO1}, \Delta H_f, \mu$	35.8	0.791	0.92
11	topostructural + topochemical + geometric + MNDO	${}^4\chi_{PC}, P_0, J, SIC_2, SIC_4, {}^5\chi^b_c, \Delta H_f$	40.4	0.765	0.97
12	topostructural + topochemical + geometric + MINDO/3	${}^4\chi_{PC}, P_0, J, SIC_2, SIC_4, E_{LUMO}$	45.8	0.758	0.98
13	topostructural + topochemical + geometric + PM3	${}^4\chi_{PC}, P_0, J, SIC_2, SIC_4, {}^5\chi^b_c, \Delta H_f$	39.7	0.761	0.98

Table V. Calculated values for the topostructural, topochemical, and AM1 quantum chemical parameters used in equation 10.

No.	χ_{PC}^4	P_0	J	SIC_2	SIC_4	χ_c^b	E_{HOMO1}	ΔH_f	μ
1	2.482	15	1.722	0.780	0.966	0.080	-9.510998	57.462489	3.246
2	1.409	10	2.356	0.824	0.875	0.059	-9.198889	-24.061979	1.613
3	1.440	11	1.993	0.831	0.975	0.058	-9.528133	51.959364	2.993
4	0.841	10	2.132	0.775	0.818	0.000	-9.761040	-22.045505	1.782
5	1.440	11	1.993	0.639	0.931	0.058	-9.342732	40.325881	1.549
6	2.209	14	1.800	0.697	0.931	0.109	-9.019172	53.561923	1.377
7	2.148	15	1.673	0.613	0.885	0.049	-8.752501	61.467301	1.686
8	3.051	17	1.694	0.616	0.890	0.119	-8.883560	90.631004	1.061
9	1.440	11	1.993	0.807	0.975	0.058	-9.497513	49.496038	1.140
10	2.650	16	1.701	0.703	0.967	0.083	-8.759018	93.256750	2.202
11	1.292	11	1.932	0.648	0.907	0.025	-8.981140	39.152911	1.625
12	3.058	17	1.692	0.593	0.890	0.112	-9.017251	86.180524	1.025
13	2.289	16	1.879	0.722	0.951	0.065	-9.635184	49.692122	5.732
14	2.154	10	2.462	0.622	0.786	0.167	-9.195396	-1.116909	1.386
15	2.136	14	1.751	0.704	0.948	0.080	-8.880375	53.383623	1.407
16	3.115	16	1.884	0.677	0.755	0.194	-9.010987	29.747467	1.402
17	1.478	9	2.346	0.719	0.867	0.083	-9.402700	5.680026	1.423
18	2.482	15	1.722	0.692	0.766	0.080	-9.008264	51.483002	0.749
19	3.131	17	1.679	0.592	0.890	0.128	-8.745169	113.597721	1.348
20	2.132	14	1.739	0.704	0.948	0.080	-9.316509	53.266008	1.795
21	2.481	16	1.832	0.699	0.902	0.103	-10.009252	50.464895	5.573
22	1.351	13	1.789	0.570	0.836	0.028	-9.611345	45.922022	1.682
23	1.418	10	2.376	0.824	0.875	0.059	-9.233259	-23.899670	2.229

24	2.132	14	1.739	0.715	0.981	0.057	-8.382162	66.295627	1.688
25	2.126	11	2.396	0.874	0.942	0.121	-10.236383	-21.118276	6.030
26	1.945	14	1.963	0.591	0.755	0.104	-8.411351	45.503434	0.270
27	2.482	15	1.722	0.791	0.967	0.080	-9.366850	8.492721	1.867
28	2.332	15	1.763	0.600	0.951	0.091	-8.782735	57.726120	1.543
29	1.478	9	2.346	0.696	0.867	0.083	-9.229828	5.699677	1.431
30	2.293	16	1.944	0.699	0.902	0.075	-9.850974	54.711440	5.793
31	1.478	9	2.346	0.847	0.910	0.083	-9.261839	-30.703134	1.260
32	2.148	15	1.673	0.651	0.891	0.049	-9.205497	91.251439	1.882
33	1.221	14	1.685	0.593	0.845	0.000	-9.510446	52.769884	1.912
34	2.499	13	2.526	0.777	0.920	0.107	-11.360524	25.435777	7.257
35	1.838	11	2.437	0.722	0.815	0.131	-8.792416	3.913795	2.561
36	1.478	9	2.346	0.836	0.962	0.083	-10.029053	-69.256743	2.575
37	1.630	15	1.681	0.603	0.659	0.000	-8.406652	39.288132	1.394
38	3.115	16	1.884	0.656	0.716	0.194	-8.782407	29.805987	2.494
39	2.913	17	1.674	0.604	0.905	0.093	-8.844299	113.962366	0.866
40	2.437	16	1.921	0.716	0.967	0.103	-9.940798	79.401262	6.265
41	3.058	17	1.700	0.616	0.920	0.119	-8.657007	101.911673	1.867
42	1.683	16	1.601	0.606	0.660	0.000	-8.707849	57.273517	2.562
43	0.816	8	2.192	0.737	0.812	0.000	-9.948850	13.095294	2.631
44	2.176	15	1.722	0.606	0.951	0.057	-8.807318	59.927756	1.359
45	0.816	8	2.192	0.737	0.812	0.000	-10.025071	-24.569648	2.776
46	2.280	15	1.787	0.603	0.885	0.091	-8.826091	57.985510	1.608
47	1.641	14	1.861	0.624	0.755	0.028	-9.637290	52.825739	0.355
48	2.888	17	1.654	0.569	0.807	0.077	-8.537199	81.775262	1.644
49	2.006	10	2.487	0.719	0.812	0.144	-9.653936	6.122184	0.948

50	2.727	18	1.612	0.786	0.920	0.080	-9.409869	19.708295	4.954
51	2.497	16	1.667	0.644	0.771	0.049	-9.614724	124.753819	2.050
52	1.292	11	1.932	0.831	0.975	0.025	-9.345759	50.639120	2.728
53	1.574	10	2.330	0.824	0.875	0.083	-9.524426	-23.745777	1.831
54	2.234	16	1.984	0.716	0.967	0.075	-9.701876	55.625683	6.167
55	1.848	14	1.867	0.628	0.902	0.066	-8.529041	45.389658	1.889
56	2.802	16	1.739	0.677	0.755	0.117	-8.724272	87.859343	1.995
57	1.683	16	1.601	0.584	0.643	0.000	-8.694071	52.783142	3.652
58	3.074	14	2.661	0.813	0.920	0.174	-11.175279	33.261219	6.162
59	1.360	12	2.246	0.740	0.890	0.059	-8.803533	-7.047410	2.543
60	1.630	15	1.681	0.579	0.642	0.000	-8.589188	21.521611	2.589
61	1.292	13	1.833	0.588	0.884	0.028	-9.075139	46.291223	1.526
62	2.802	16	1.744	0.677	0.771	0.117	-8.760423	87.878976	2.958
63	2.293	14	1.786	0.697	0.931	0.127	-8.809819	52.914796	1.658
64	2.972	17	1.656	0.613	0.896	0.093	-8.672342	86.560420	1.569
65	1.138	8	2.279	0.775	0.962	0.083	-9.647217	13.148070	1.773
66	2.214	11	2.461	0.788	0.903	0.250	-10.328717	-135.798912	4.070
67	2.274	14	2.092	0.732	0.939	0.093	-9.498965	42.132738	5.212
68	2.332	16	1.793	0.699	0.902	0.065	-9.707684	49.439690	6.645
69	0.816	8	2.192	0.737	0.812	0.000	-9.958995	24.673699	2.834
70	1.478	9	2.346	0.885	0.966	0.083	-9.512320	-30.257131	1.873
71	2.994	18	1.913	0.670	0.725	0.146	-8.597273	-29.701343	0.593
72	1.351	13	1.789	0.633	0.783	0.048	-9.618662	-11.036978	1.453
73	1.221	14	1.685	0.593	0.845	0.000	-9.519593	24.038959	3.243
74	2.855	19	1.809	0.670	0.738	0.118	-8.322206	14.345758	1.347
75	3.130	17	1.674	0.786	0.953	0.117	-9.907587	57.088597	7.715

76	1.759	14	1.780	0.558	0.624	0.028	-8.898246	44.312986	2.417
77	2.390	14	2.079	0.760	0.939	0.103	-9.995923	44.945430	7.318
78	2.348	16	1.843	0.699	0.902	0.065	-10.065351	48.997787	5.907
79	2.391	16	1.760	0.656	0.836	0.065	-10.153390	48.597189	7.636
80	2.300	15	1.714	0.655	0.884	0.083	-9.466774	90.375028	1.894
81	2.975	17	1.775	0.705	0.773	0.167	-8.668864	-51.583170	2.233
82	1.851	11	2.471	0.863	0.938	0.070	-10.795945	14.958329	5.163
83	1.292	11	1.932	0.807	0.975	0.025	-9.250508	61.289442	2.564
84	2.136	14	1.751	0.715	0.981	0.057	-8.650669	70.561209	2.432
85	1.478	9	2.346	0.738	0.875	0.083	-9.338439	12.337686	1.935
86	2.180	15	1.741	0.606	0.935	0.057	-8.832492	56.103853	1.663
87	1.700	14	1.820	0.611	0.869	0.028	-8.581538	44.585899	2.808
88	2.300	15	1.714	0.617	0.896	0.083	-9.168383	66.520403	1.216
89	2.293	14	1.786	0.708	0.962	0.091	-8.617125	69.956608	1.276
90	2.357	15	1.760	0.587	0.787	0.103	-9.179235	64.230081	1.689
91	2.209	14	1.800	0.708	0.962	0.082	-8.497152	66.236222	1.211
92	3.175	19	1.575	0.553	0.913	0.124	-8.830777	100.875189	1.130
93	3.110	17	1.677	0.577	0.890	0.112	-8.958369	70.826740	1.287
94	3.721	21	1.867	0.638	0.674	0.263	-8.315255	10.633206	1.225
95	2.497	16	1.664	0.644	0.755	0.049	-9.634497	124.742897	0.004

On the Relationship between the Organic-Carbon Normalized Sediment, or Soil, Sorption Coefficient and the Octanol-Water Partition Coefficient

K.B. Lodge

Department of Chemical Engineering, University of Minnesota, 10 University Drive, Duluth, Minnesota 55812-2496

Abstract. If the organic carbon in sediment or soil has the same partitioning properties as octanol, the relationship between the organic-carbon normalized sorption coefficient, K_{oc} with units of L/Kg, and the octanol-water partition coefficient, K_{ow} , is $\log K_{oc} = \log K_{ow} + 0.214$ at 20 °C. Observations are well represented by $\log K_{oc} = \log K_{ow} - 0.289$; this is calculated using the data critically reviewed by Baker and coworkers (Water Environ. Res., 69(2), p136-145 (1997)). We conclude that partitioning properties of the organic carbon in sediment or soil are not the same as octanol; experimental values of K_{oc} are about one third of those expected if the organic carbon behaves identically to octanol.

When considering the distribution of hydrophobic nonionic organic compounds in aqueous systems, are the partitioning properties of octanol the same as those of the organic carbon in soils or sediments? If so, we may be tempted to suppose that

$$K_{oc} \approx K_{ow} \quad [1]$$

where K_{oc} is the sediment, or soil, sorption coefficient and K_{ow} is the octanol-water partition coefficient. These coefficients are defined as follows:

$$\begin{aligned} K_{oc} &= C_s / (F_{oc} \cdot C_{aq}) \\ K_{ow} &= C_{oct} / C_{aq} \end{aligned} \quad [2]$$

where C_s is the mass of chemical per unit mass of dry sediment or soil, F_{oc} is the fraction of organic carbon in the dry sediment or soil, C_{aq} is the mass of chemical per unit volume of aqueous phase, and C_{oct} is the mass of chemical per unit volume of octanol. There is ample evidence to believe that eq 1 holds within an order of magnitude or so. Because the partitioning of nonionic organic chemicals is a purely physical process, we may be tempted to believe that the organic carbon in the sediment or soil is no different from that in octanol and so it is reasonable to suppose that eq 1 holds.

Our purpose is to put eq 1 on a more formal footing and examine closely the ideas behind it. We do not like the equation as it stands unless we clearly recognize it to be a dimensional equation (units of K_{oc} are L/Kg here and throughout); the concentration basis on the left-hand

side is not the same as the right-hand side. The concentration of the chemical in the sediment, C_s , is defined as the mass of chemical per *unit mass of dry sediment or soil* and the concentration of the chemical in the octanol is defined as the mass chemical per *unit volume of octanol*. However, we may develop a relationship between K_{oc} and K_{ow} , ensuring a consistent concentration basis thereby.

To do this, we take the idea that the organic carbon in the sediment is no different from that in the octanol and construct a model for which we consider the equilibrium. The model consists of sediment, water and octanol in a container (see figure 1a) and we consider the distribution of a chemical at equilibrium within the container. The octanol, being the least dense phase, is uppermost. Now we assume that the distribution behavior is explained by considering the sediment to behave like octanol. Practically, we would have to prepare the system in the configuration as shown in figure 1b because octanol is less dense than water.

At equilibrium

$$C_{aq} = C_{oct}^o / K_{ow}^o = C_{oct}^s / K_{ow}^s$$

or

$$K_{ow}^o = C_{oct}^o / C_{aq} = K_{ow}^s = C_{oct}^s / C_{aq} \quad [3]$$

where the superscripts "o" and "s" represent the octanol and the octanol phase representing the sediment respectively. Now we consider the concentration of the chemical in the octanol phase that represents the sediment.

$$C_{oct}^s = m / V_{oct} \quad [4]$$

where m is the mass of chemical in the volume of octanol, V_{oct} . Now we express this concentration in terms of the mass of chemical per *unit mass of octanol*.

$$C_{oct}^s = m \cdot \rho_{oct} / M_{oct} \quad [5]$$

where M_{oct} is the mass of octanol and ρ_{oct} is the density of octanol. This can be written as

$$C_{oct}^s = C_s^s \cdot \rho_{oct} = (C_s^s / F_{oc}^{oct}) \cdot F_{oc}^{oct} \cdot \rho_{oct}$$

where F_{oc}^{oct} is the fraction of organic carbon in octanol. Now, from eq 3,

$$K_{ow}^o = C_{oct}^s / C_{aq} = [C_s^s / (F_{oc}^{oct} \cdot C_{aq})] \cdot F_{oc}^{oct} \cdot \rho_{oct}$$

The concentration term within the square brackets on the right-hand side is just K_{oc}^s , the organic-carbon normalized sorption coefficient for the octanol phase representing the sediment. So

$$K_{ow}^o = K_{oc}^s \cdot F_{oc}^{oct} \cdot \rho_{oct}$$

The units are now consistent. In the logarithmic form

$$\log K_{oc} = \log K_{ow} - \log(F_{oc}^{oct} \cdot \rho_{oct}) \quad [6]$$

Here we have dropped the superscripts because the two phases are indistinguishable. The relationship is weakly dependent upon temperature because of the density term. We calculate the fraction of organic carbon in octanol from the relative molecular masses; $F_{oc}^{oct} = 0.738$; the density of octanol (1) at 20 °C is 0.827 g/mL. So, at 20 °C,

$$\log K_{oc} = \log K_{ow} + 0.214 \quad [7]$$

If the organic carbon in the sediment behaves identically to octanol, then we expect the relationships in eq 6 and 7 to be observed. What is observed? From experimental data, many workers have developed relationships of the general form

$$\log K_{oc} = a \cdot \log K_{ow} + b \quad [1]$$

where values of a and b are determined by linear regression. To make the essential point here, we consider only the recent relationship developed by Baker and coworkers(2); they developed selection criteria and critically reviewed the available measurements. They found the following values, for $1.7 < \log K_{ow} < 7.0$, using data for 72 chemicals:

$$a = 0.903 \pm 0.034; \quad b = 0.094 \pm 0.142; \quad \text{and} \quad r^2 = 0.91$$

We wish to compare eq 7 with this result.

However, the equilibrium model that we used requires $a = 1$; other values of a do not have a physical meaning for the model considered here. Using the data of Baker and coworkers, we applied a regression model in which a is forced to be unity(3). We obtain

$$\text{for } a = 1; \quad b = -0.289; \quad \text{and} \quad r^2 = 0.90$$

If octanol and the organic carbon in the sediment or soil are indistinguishable, then we expect $b = +0.214$ (see eq 7). In other words, experimental values of K_{oc} are about one third of the values expected if the sediment or soil organic carbon were to have the same partitioning properties as octanol. Whereas eq 1 is a useful first approximation, it should not be concluded therefrom that the partitioning properties of octanol and the organic carbon in sediment or soils are the same.

Acknowledgments. We acknowledge support provided by the U.S. EPA with cooperative agreements CR-813504 and CR-817486, and the U.S. Air force with award number F49620-94-1-0401. The ideas presented here were developed during experimental work done under these grants; they should in no way be taken to represent the opinions of the grantors.

References

1. Daubert, T.E., "Physical and Thermodynamic Properties of Pure Chemicals: Data Compilation", Hemisphere Publishing Corporation, New York, 1989.
2. Baker, J.R., J.R. Mihelcic, D.C. Luehrs & J.P. Hickey, "Evaluation of Estimation Methods for Organic Carbon Normalized Sorption Coefficients", Water Environ. Res., **69**, 136 (1977).
3. KaleidaGraph 3.07, Synergy Software, Reading, Pennsylvania, July 5, 1996.

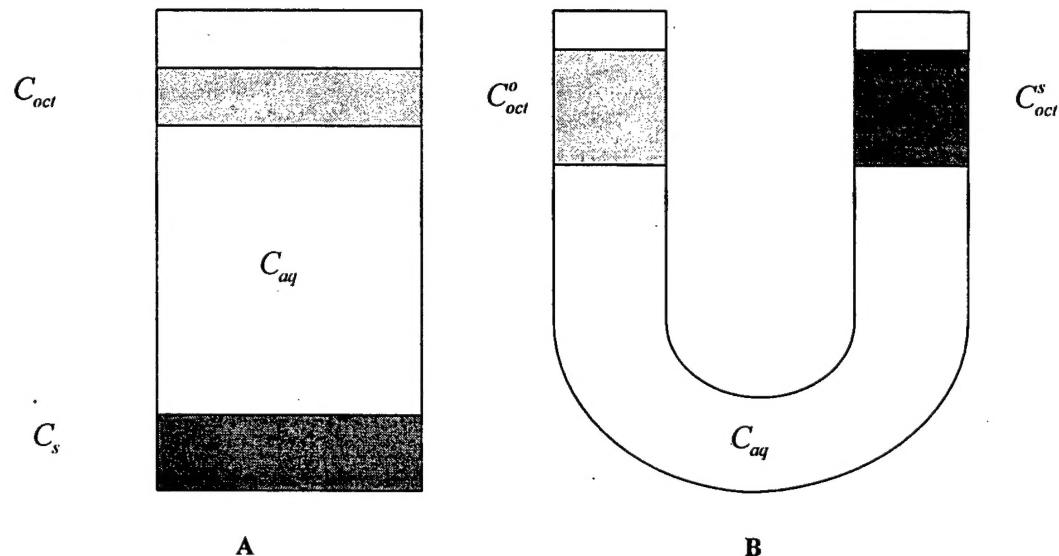


Figure. This shows the model used to examine the relationship between K_{ow} and K_{oc} . In A octanol is placed on a aqueous phase that lies over the soil or sediment. Upon assuming that the sediment behaves like octanol, the system would adopt the configuration shown in B. The model considers the distribution of the chemical between the phases at equilibrium; the symbols, representing the concentrations of chemical in the various phases, are defined in the text.

Approved for public release,
distribution unlimited

AIR FORCE
NOTICE
This tec
approver
Distributor
Joan Geiger
STINFO Program Manager

150

15-1061
PUB PERIOD